COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 47 (2023) e13395

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13395

Relevance and the Role of Labels in Categorization

Felix Gervits,[a] Megan Johanson,[b,c] Anna Papafragou[d]

[a]*Department of Linguistics and Cognitive Science, University of Delaware*
[b]*Department of Psychological and Brain Sciences, University of Delaware*
[c]*Practice Analytics Team, Mayo Clinic, Mankato*
[d]*Department of Linguistics, University of Pennsylvania*

Received 2 November 2021; received in revised form 20 November 2023; accepted 27 November 2023

Abstract

Language has been shown to influence the ability to form categories. Nevertheless, in most prior work, the effects of language could have been bolstered by the fact that linguistic labels were introduced by the experimenter prior to the categorization task in ways that could have highlighted their relevance for the task. Here, we compared the potency of labels to that of other non-linguistic cues on how people categorized novel, perceptually ambiguous natural kinds (e.g., flowers or birds). Importantly, we varied whether these cues were explicitly presented as relevant to the categorization task. In Experiment 1, we compared labels, numbers, and symbols: One group of participants was told to pay attention to these cues because they would be helpful (Relevant condition), a second group was told that the cues were irrelevant and should be ignored (Irrelevant condition), and a third group was told nothing about the cues (Neutral condition). Even though task relevance affected overall reliance on cues during categorization, participants were more likely to use labels to determine category boundaries, compared to numbers or symbols. In Experiments 2 and 3, we replicated and fine-tuned the advantage of labels in more stringent categorization tasks. These results offer novel evidence for the position that labels offer unique indications of category membership, compared to non-linguistic cues.

Keywords: Categorization; Relevance; Labels; Language and thought

1. Introduction

It is widely assumed that the way people categorize objects benefits from information about how the objects are named. For instance, when two objects are given the same label, people tend to think of them as more similar to each other, compared to when objects do not share

Correspondence should be sent to Felix Gervits, US Army Research Laboratory, ARL Northeast, 141 S Bedford St., Burlington, MA 01803, USA. E-mail: felix.gervits.civ@army.mil

a label (Goldstone, 1994). People recall (Carmichael, Hogan, & Walter, 1932) and categorize (Fairchild & Papafragou, 2019; Johanson & Papafragou, 2016) perceptually ambiguous stimuli in accordance with the way the stimuli have been labeled. Moreover, people rely more on labels than on perceptual features in making category decisions when the two cues conflict (Deng & Sloutsky, 2012; Gelman & Davidson, 2013; Sloutsky, Lo, & Fisher, 2001). It has also been shown that linguistic labels influence the way people learn novel categories even when the labels are completely redundant (Lupyan, Rakison, & McClelland, 2007; see also Fisher, 2010; Hoffman & Rehder, 2010; Lupyan & Thompson-Schill, 2012; Markman & Ross, 2003; Yamauchi, Kohn, & Yu, 2007; Yamauchi & Markman, 1998, 2000; Yamauchi & Yu, 2008). The influence of labels on categorization appears early in development (Booth & Waxman, 2002a; Casasola & Bhagwat, 2007; Ünal & Papafragou, 2016; Waxman & Markow, 1995). For instance, even young children expect that objects that look similar should share the same label (Gelman & Markman, 1986) and treat shared labels as indicating similarities between objects (Althaus & Plunkett, 2015; Balaban & Waxman, 1997; Booth & Waxman, 2002b; Fairchild, Mathis, & Papafragou, 2018; Fulkerson & Waxman, 2007; Perszyk & Waxman, 2018; Welder & Graham, 2006). Furthermore, children can use labels to categorize novel exemplars even when the labels conflict with perceptual appearance (Gelman & Markman, 1986; Johanson & Papafragou, 2016; Landau & Shipley, 2001; Plunkett, Hu, & Cohen, 2008; but see Deng & Sloutsky, 2012; Gelman & Davidson, 2013, for some limitations).

Together these data have been interpreted as support for the position that labels function as category markers for both adults and children (Balaban & Waxman, 1997; Gelman & Davidson, 2013; Gelman & Markman, 1986). On this position, labels function as an invitation to form categories (Waxman & Markow, 1995): "[e]xactly what makes a dog a dog, or a lamb a lamb, may be unknown […], but a category label can serve as a placeholder that a reason exists" (Jaswal & Markman, 2007, p. 96). According to this view, by helping identify two objects as members of the same category, labels establish their equivalence, making it possible to recognize new members of the category and to make inferences about non-obvious properties from one member of the category to another. These steps can have cascading cognitive consequences by highlighting certain conceptual distinctions over others (Lupyan, 2008), enabling further linguistic and conceptual learning (Ferguson & Waxman, 2016), and allowing named categories to be socially shared, thereby promoting the alignment of conceptual representations with others within a community (Boyd, Richerson, & Henrich, 2011). Nevertheless, debate continues about the precise mechanism underlying the role of labels in supporting object kind conjectures (see Anderson, 1991; Deng & Sloutsky, 2012; Gliozzi, Mayor, Hu, & Plunkett, 2009; Perfors & Navarro, 2010), and several issues remain about the potency of labels as categorization cues.

Notice that, in virtually all previous studies with adults (and many in children) that have pointed to a strong role for language in categorization, linguistic labels were introduced overtly by the experimenter. Specifically, labels were presented alongside a newly introduced exemplar of a category ("This is an X" or simply "X") and were often explicitly linked to a kind, not to the individual exemplar itself. For instance, Yamauchi and Markman (2000) asked adults to make classification and inference judgments on novel stimuli based on nonsense labels (e.g., "monek," "plaple") and told participants that "the labels …represented

two types of imaginary bugs." Deng and Sloutsky (2012) provided people with either an auditory or a written novel label (e.g., "flurp"), and instructed them that these labels represented novel objects (see also Sloutsky et al., 2001). Similarly, in a study showing that adults learned to categorize "aliens" faster after learning novel labels for them, compared to no labels, the experimenters "encouraged subjects to think of the labels as referring to kinds, rather than properties" (Lupyan et al., 2007; see also Johanson & Papafragou, 2016; Lupyan & Thompson-Schill, 2012, Experiment 4). Since in these studies, labels were presented as part of the task instructions, it is perhaps not surprising that adults took the labels to be relevant to the task. After all, communication in general carries a presumption of relevance (Sperber & Wilson, 1986), and it would be reasonable to assume that the fact that the stimuli in these studies were overtly named (and sometimes explicitly linked to kinds) suggested that they must have been meant by the experimenter to have some bearing on the task at hand. In contrast to the view that labels serve as category markers, this opens up a new interpretation of prior results. Specifically, participants in such categorization tasks might have developed a strategic bias to respond to the expectation that items that have just been labeled alike should belong to the same category without actually integrating the information from labels with perceptual and other information about the categorical structure of the stimuli (Goldstone, Lippa, & Shiffrin, 2001). A more stringent test of the theoretical position that labels are interpreted as category markers would involve dissociating the information carried by the labels from the act of labeling the stimuli during the categorization task (or any other instructions that might suggest that labels are relevant for solving the task).

Relatedly, some of the studies described above have reported that linguistic labels are privileged as categorization cues, compared to non-linguistic cues. In the work by Lupyan et al. (2007), adults learned to categorize "aliens" faster after learning novel labels for them, compared to seeing behaviors that indicated where the aliens lived (see also Lupyan & Thompson-Schill, 2012). Additionally, Balaban and Waxman (1997) demonstrated that 9-month-old children formed categories for objects when given an associated word but not when given an auditory tone as a cue (cf. Fulkerson & Haaf, 2003). Notice, however, that in these studies, labels were the only cues that were directly communicated (uttered) by the experimenter and hence presumably identified as relevant. Non-label cues were instead accessed through exposure to experimenter-independent eventualities (the aliens' movements in Lupyan et al., 2007; tones played in isolation in Balaban & Waxman, 1997) and may not have carried the same presumption of relevance (in the sense of Sperber & Wilson, 1986). Especially for adults, it is therefore important to compare labels to other symbolic but non-linguistic cues *under the same task-relevance conditions* to ascertain that labels produce unique and powerful effects on categorization.

In this paper, we present the first study to examine the role of relevance on adults' reliance on labels (and other cues) during categorization. In three experiments, we adopt the same basic paradigm, in which adults had to categorize novel natural kind exemplars (e.g., novel flowers or birds) that were perceptually equidistant from two Standards (see Johanson & Papafragou, 2016). In these cases, perceptual information gave no indication of category membership. We asked whether a shared *novel* label (noun) might motivate adults to group such perceptually ambiguous stimuli with one of the two Standards. We also compared the

usefulness of labels to other cues such as numbers and symbols that do not refer to object kinds. We used numbers because, like labels, they were meaningful. We used arbitrary symbols (rotated numbers) because, unlike labels and numbers, they had little or no meaning (even though in terms of low-level perceptual features they were identical to the numbers). The use of non-label cues enables us to compare to prior work, which has investigated the effect of various cues on categorization and found a label advantage (e.g., Johanson & Papafragou, 2016). A crucial difference from previous studies is that the cues were presented in written, not auditory, format and were never introduced overtly by the experimenter (e.g., labels never appeared in sentences such as "This is an X"): In fact, all cues were dissociated from an overt communicator altogether. This design allowed us to manipulate global information about the purposefulness or relevance of the cues given to participants at the beginning of the experiment, such that cues were presented as Relevant, Irrelevant, or Neutral (unspecified) for the main categorization task.

The present design allows us to address specific—but currently untested—hypotheses about the role of language in categorization, especially in the absence of convergent perceptual cues about the structure of a category. We expect that, overall, participants' use of all cues should be sensitive to the relevance manipulation (Campbell & Namy, 2003; Fulkerson & Haaf, 2003; Jaswal, 2004). For instance, alongside labels, it is possible that cues such as numbers or symbols would also function as an "invitation to form categories" when explicitly introduced as task-relevant but would be ignored if presented as irrelevant. In the absence of specific instructions (cf. the Neutral condition), both labels and non-label cues might shape categorization as long as participants are free to assume that such cues should be taken as potentially relevant (Sperber & Wilson, 1986). Crucially, however, on the position that labels are category markers, adults' reliance on non-label cues for category formation should be more limited compared to labels across relevance manipulations: Unlike labels, numbers and symbols are not *referentially* linked to the natural kind categories in the experimental stimuli and, as a result, should not support kind-membership with the same precision as labels. Thus, beyond the general expectation that the task relevance of a cue should matter, this account makes the unique prediction that novel nominal labels should guide categorization more strongly than numbers or symbols, other things being equal, *even when task relevance is otherwise held constant*.

2. Experiment 1

2.1. Methods

2.1.1. Participants

A total of 155 English-speaking adults were used as participants in the study. They were recruited from Introductory Psychology courses at the University of Delaware and were given partial course credit for participating.

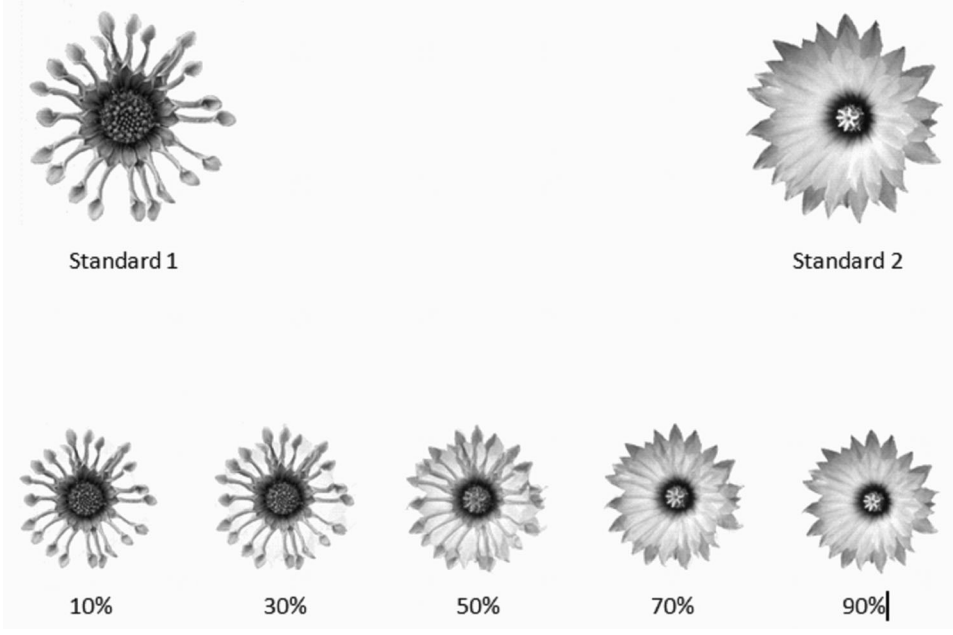

Fig. 1. A sample set of Standards and Targets.

2.1.2. Stimuli and procedure

Four sets of black-and-white stimuli were created: one set of birds, two different sets of flowers, and one set of fish. Each of these sets consisted of photographs of two stimuli (Standards) that were morphed through FantaMorph (a commercial image morphing program) into five different stimuli (Targets) along a scale of varying similarity to the originals (see also Johanson & Papafragou, 2016). Each Target was 10%, 30%, 50%, 70% or 90% similar to one of the Standards as determined by FantaMorph (see Fig. 1).[1] The 10% and 30% Targets were always perceptually closer to Standard 1, whereas the 70% and 90% Targets were always perceptually closer to Standard 2. The 50% Target was perceptually ambiguous, that is, equidistant from the Standards. To confirm these rankings, a separate group of 13 adults was presented with triads consisting of the two Standards and each of the Targets and was asked which of the Standards each Target "went with." People in this group were highly accurate with the 10%, 30%, 70%, and 90% (unambiguous) trials ($M = 0.96$) but were at chance with the 50% (ambiguous) trials ($M = 0.58$; $t(12) = 17.73$, $p < .001$).

Within each of the four stimuli sets, there were five trials for a total of 20 trials. Each trial included a triad display, with the two Standards on top and a morphed Target image (the 10%, 30%, 50%, 70%, or 90% Target) at the bottom, separated from the Standards by a solid line. This triad display was presented for 8 s (study phase). Then the objects disappeared for 500 ms, but the solid line remained. The objects reappeared with a red border surrounding the display and stayed on the screen for 7 s (test phase). All five trials within a set were presented in block sequence. The fourth trial in each set was always the ambiguous (50%)

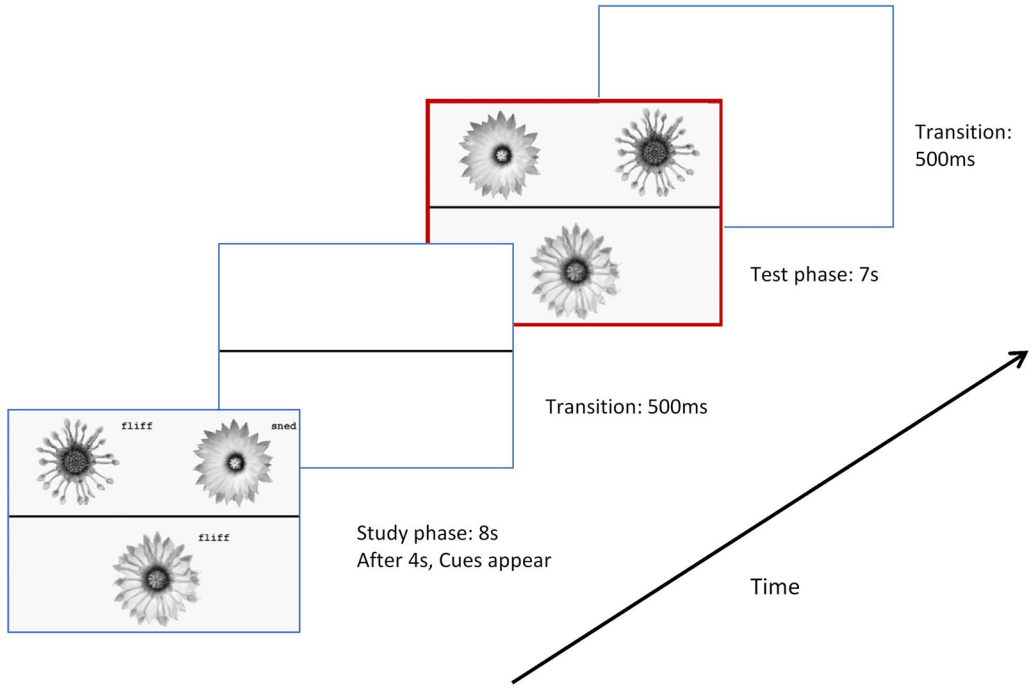

Fig. 2. Sample ambiguous trial in Experiment 1.

trial, but the order of the unambiguous trials varied randomly for each set. Blank screens marked transitions between trials within a set (500 ms) and between sets (2 s). For half of the participants, the left-right position of the Standards was switched in the test phase so as to prevent any side associations.

Within each set, visual cues appeared 4 s (halfway) into the study phase for the ambiguous trial and for one additional, randomly selected trial that came before the ambiguous one (Cue trials). Three cues appeared simultaneously within each display, with each one being matched to one of the natural kind exemplars (Standards and Target) in a triad. Each cue would appear in the top-right corner of each exemplar, blink twice, and then stay on for another 2 s until the end of the study phase for that trial (see Fig. 2 for a sample trial). The cue for the Target was identical to the cue for one of the Standards. In the ambiguous Cue trials, the cue for the Target arbitrarily matched that of either the left or right Standard. In the unambiguous Cue trials, the Target cue always respected the perceptual similarity between Target and Standards (these trials were included so that the ambiguous Cue trials would not be the only ones accompanied by Cues—a feature that might have encouraged trial-specific strategies). The assignment of cues to Standards and Target was counterbalanced across participants.

There were three types of Cues: Labels, Numbers, and Symbols. Participants were randomly assigned to one of three groups depending on Cue type. The labels were “lorp” and “pim” (bird set), “fliff” and “sned” (first flower set), “blick” and dax (second flower set), “hep” and “moof” (fish set). The numbers were 6 and 3, 1 and 5, 2 and 7, and 8 and 4 for the

Table 1
Sample cues for Experiment 1

Cue	Example	
Label	lorp	pim
Number	3	6
Symbol	ε	ρ

corresponding sets, presented in a distinctive font. The symbols were the number cues rotated 90 degrees clockwise and hence were identical to numbers in terms of low-level visual properties (Lupyan & Spivey, 2008). See Table 1 for examples of cues.

Participants were tested in small groups of five to seven people seated in front of a projector screen, which displayed the stimuli. Within each Cue group, participants were randomly assigned to one of three Relevance conditions (Neutral, Irrelevant, Relevant) depending on the instructions given to them at the beginning of the session. Instructions were read aloud to all participants. In the Neutral condition, the instructions were as follows: "In this experiment, you will be presented with a series of slides, each containing three images. Your task is to match the bottom image as best you can with one of the top two images. Each slide will appear twice. The first time, you will have 8 s to inspect the slide. The slide will briefly disappear and reappear again with a red border around it. Your task then will be to mark down on your answer sheet, which of the top two images the bottom image on the slide best goes with. Mark L if the bottom image matches the top left image, or mark R if the bottom image matches the top right image. Pay attention as the position of the top two images may have been switched around between the first and second time you see the slide. Please only write your answers when you see the red border around the slide."

In the other two conditions, there was additional information at the end of these instructions. People in the Irrelevant condition were told: "We have been having issues with our software this week, so occasionally you might see random messages displayed on the screen. These are actually a glitch from another experiment, so please disregard them as they are irrelevant to your task." People in the Relevant condition were told: "Pay attention to all information on the screen as it will be helpful in your task." Participants marked their responses sequentially on an answer sheet. Overall, the study took about 20 min to complete.

2.2. Results

We first examined the percentage of correct responses for unambiguous trials (10%, 30%, 70%, and 90%). In these trials, correct responses were those that conformed to perceptual similarity. As expected, performance was highly accurate on the three unambiguous trials

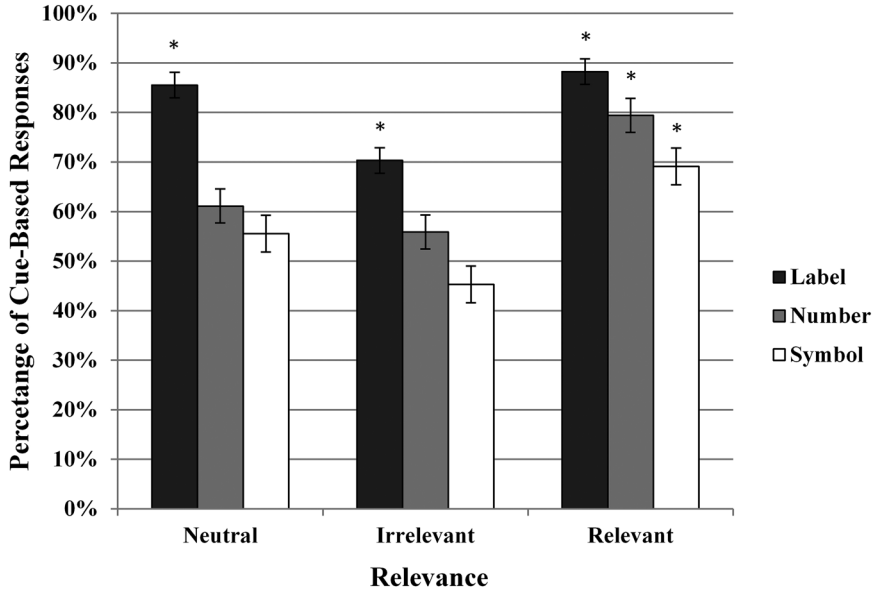

Fig. 3. Percentage of Cue-based Responses for ambiguous trials in Experiment 1. Asterisks reflect significant differences from chance (50%) performance ($p < .05$). Error bars represent standard error of the mean (SEM).

per set that were presented without cues in all Cue ($M_L{}^2 = 1.00$, $M_N = 0.98$, $M_S = 0.97$) and Relevance subgroups of participants ($M_N = 0.99$, $M_I = 0.98$, $M_R = 0.98$). Similarly, for the one unambiguous trial per set that was also presented alongside cues, performance was at the ceiling across Cue ($M_L = 1.00$, $M_N = 0.99$, $M_S = 0.97$) and Relevance subgroups of participants ($M_N = 1.00$, $M_I = 0.99$, $M_R = 0.97$). Thus, participants were successful in categorizing the unambiguous stimuli with or without additional cues.

For ambiguous (50%) trials, the data for Experiment 1 consisted of 155 participants × 4 items = 620 observations. Fig. 3 summarizes the data. Data were analyzed using multilevel mixed-effects logistic regression. We used a model that included Cue-based Responses (i.e., responses that followed the cue given) as the binary dependent variable and participants as a random intercept. The addition of items as a random intercept did not improve the model fit. Relevance was Helmert-coded with Neutral versus Irrelevant as the first comparison and Relevant versus not-Relevant (the average of Neutral and Irrelevant) as the second comparison. Cue was also Helmert-coded with Symbol versus Number as the first comparison and Label versus not-Label (the average of Symbol and Number) as the second comparison. The best fit for these data was a model that included the main effects of Cue and Relevance. The cross-level interaction between Cue and Relevance did not significantly improve model fit and was, therefore, not included in the final model. Table 2 presents a regression table for the model of Cue-based Responses for Experiment 1.

The model revealed a significant effect of Cue (Label vs. not-Label): There were more Cue-based Responses for ambiguous trials presented with Labels ($M_L = 0.82$, $\beta = 1.075$, $SE = 0.210$, $z = 5.123$, $p < .001$) than with other cues. However, there was no significant effect

Table 2
Multilevel mixed-effects logistic regression model of Cue-based Responses for Experiment 1

Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	z	p
Intercept	0.0837	0.0932	8.933	4.13e-16***
Cue (Symbol vs. Number)	−0.3863	0.2077	−1.860	.062916
Cue (Label vs. not-Label)	1.0749	0.2098	5.123	3.01e-07***
Relevance (Neutral vs. Irrelevant)	0.4658	0.2103	2.215	.026770*
Relevance (Relevant vs. not-Relevant)	0.8517	0.2047	4.161	.000032***

Note. *$p < .05$; **$p < .01$; ***$p < .001$.

of Cue (Symbol vs. Number): There were no significant differences between Symbol ($M_S = 0.57$) and Number ($M_N = 0.65$, $\beta = -0.386$, $SE = 0.208$, $z = -1.860$, $p = .063$). For the Relevance main effect, both terms were significant. Participants were less reliant on the cue in the Irrelevant condition, compared to the Neutral condition ($\beta = 0.466$, $SE = 0.210$, $z = 2.215$, $p = .0268$) and more reliant on the cue in the Relevant condition, compared to the Irrelevant and Neutral conditions together ($\beta = 0.852$, $SE = 0.205$, $z = 4.161$, $p < .001$).

As Fig. 3 shows, performance with Labels was significantly different from chance in all conditions (Neutral: $M_L = 0.86$, $t(18) = 10.21$, $p < .001$; Irrelevant: $M_L = 0.70$, $t(15) = 3.90$, $p < .01$; Relevant: $M_L = 0.88$, $t(16) = 10.10$, $p < .001$). By contrast, performance with Numbers differed from chance in the Relevant condition ($M_N = 0.79$, $t(16) = 6.00$, $p < .001$), but was no different from chance in the Neutral ($M_N = 0.61$, $t(18) = 2.05$, $p = .057$) and Irrelevant condition ($M_N = 0.56$, $t(16) = 0.94$, $p = .361$). Finally, performance with Symbols differed from chance only in the Relevant condition ($M_S = 0.69$, $t(16) = 3.79$, $p < .01$) but was at chance levels in both the Neutral ($M_S = 0.56$, $t(18) = 0.85$, $p = .409$) and the Irrelevant condition ($M_S = 0.45$, $t(15) = -0.72$, $p = .485$).

2.3. Discussion

In Experiment 1, we found that labels were particularly potent in influencing category decisions about perceptually indeterminate exemplars of novel natural kinds when compared to other cues such as numbers and symbols. Furthermore, across cues, explicitly stated task relevance shifted reliance on cues, compared to the remaining cases as a whole. Similarly, explicit indications of task irrelevance affected people's willingness to follow the cues, compared to a Neutral control. A particularly interesting finding was that people used novel labels to constrain their categorization decisions at levels different from chance even when the labels were not tied to a specific intent (Neutral condition). Perhaps surprisingly, this effect carried over even to cases where labels were presented as the result of an accident (Irrelevant condition). By contrast, numbers and symbols were clearly used at above-chance levels only when their relevance to the task was explicitly highlighted (Relevant condition).

Unlike past experiments, labels here were presented visually, were not introduced overtly as labels, and were not even identified as linguistic stimuli. Could it be that participants did not interpret the novel letter strings in Experiment 1 as labels? To test this hypothesis, we devised familiar labels for all stimuli (e.g., "carnation" and "marigold" for Standards 1 and 2 in Fig. 1).

To ensure that these labels were appropriate for our stimuli, we asked 40 participants to rate each Familiar Label on a scale from 1 (*least matching*) to 7 (*perfect match*) in terms of "how accurately the words match their respective images" (i.e., Standards). The average rating was 4.76 ($SD = 1.77$), suggesting that people generally found these labels to represent the stimuli. In a control manipulation, we then replicated the Label condition from Experiment 1 with 60 new participants from the same pool but replaced the novel labels with Familiar Labels. Overall, the results replicated the Label results in Experiment 1. An analysis comparing cue-compliance for the 50% trials with Label Type (Experiment 1 Label, Control Familiar Label) and Relevance as factors yielded only a significant effect of Relevance ($p = .013$), but no significant effect of Label Type ($p = .936$), and no interaction ($p = .546$). These findings support our hypothesis that participants relied on novel labels to categorize unfamiliar natural kinds in Experiment 1 because they truly took these labels to refer to the newly introduced categories.

The findings from Experiment 1 provide novel support for the position that labels indicate category membership. First, as this position expects, adults' use of labels is driven by their perceived relevance since this manipulation is tied to the goals of the task. Nevertheless, the power of labels persists even in conditions in which the intention to direct people's attention to the labels and indicate their relevance to the task is weak (Neutral condition) or absent (Irrelevant condition). We return to this finding in later experiments. Second, as this position predicts, even though an explicit invitation to consider non-linguistic cues such as numbers and symbols as relevant does affect their use to guide categorization, adults' reliance on labels exceeds their reliance on other cues across identical task-relevance manipulations.

3. Experiment 2

The results from Experiment 1 raise two issues. First, one might be concerned that the task simply instructed participants to find a visual "match" to the Target image and did not clearly point to categorization. In Experiment 2, we therefore replicated the basic design of Experiment 1 but modified the instructions subtly to clarify the intent of the task. We only included two types of cues, numbers and labels since we found no difference between numbers and symbols in Experiment 1. Second, our current data leave open the possibility that people were simply not attending to the non-label cues. To address this possibility, at the end of Experiment 2, we conducted a surprise memory check to measure how well people remembered the cues previously given. Evidence that people recalled the cue types similarly would cast doubt on the role of cue salience in the categorization results.

3.1. Methods

3.1.1. Participants

A total of 120 English-speaking adults participated in the study. They were recruited from Introductory Psychology courses at the University of Delaware and were given partial course credit for participating. None of them had participated in the earlier study.

3.1.2. Stimuli and procedure

The stimuli and procedure were the same as in Experiment 1 (Label and Number conditions) except for a change to the task instructions to more clearly reflect a categorization task. Specifically, we replaced the Experiment 1 sentence: "Your task is to match the bottom image as best you can with one of the top two images" with "Your task is to determine as best you can which of the top two images the bottom image best goes with." This phrasing has been used in many past studies on labels and categorization (e.g., Johanson & Papafragou, 2016).

We also introduced a surprise memory check at the end of the task to gauge how well participants could recall the cues. All participants were presented with a grid of 48 cues on an 8.5 × 11-inch sheet of paper. We included the main cues of Label ($N = 8$) and Number ($N = 7$, one number was accidentally omitted from the memory task). We also included eight distractor cues for Numbers and (because of an error) nine distractor cues for Labels. These within-category distractors were simply different nonsense words or numbers presented in the same font and size as the original stimuli. In addition, a total of 16 novel shapes were included as additional (outside-category) distractors. Participants were told: "Please circle all the messages that you saw during the course of the experiment." The study took about 25 min to complete.

3.2. Results

3.2.1. Categorization task

We examined the percentage of correct (i.e., perception-driven) responses for the unambiguous trials that lacked cues (10%, 30%, 70%, and 90%). The results were near-ceiling for all Cue ($M_L = 0.95$, $M_N = 0.96$) and Relevance subgroups of participants ($M_N = 0.97$ $M_I = 0.97$, $M_R = 0.96$). For unambiguous trials that were accompanied by cues, participants were near-ceiling for all Cue ($M_L = 0.95$, $M_N = 0.98$) and Relevance manipulations ($M_N = 0.98$ $M_I = 0.97$, $M_R = 0.96$). Thus, participants were able to categorize the unambiguous stimuli with or without additional cues, much like in Experiment 1.

For ambiguous (50%) trials, the analysis dataset for Experiment 2 consisted of 120 participants × 4 items = 480 observations. Fig. 4 summarizes the data. The same data analytic strategy as in Experiment 1 was used. Our analysis used a model that included Cue-based Responses to ambiguous trials as the binary dependent variable, Cue (Label, Number) and Relevance (Neutral, Irrelevant, Relevant) as fixed predictors, and a random intercept for participants. The addition of items as a random intercept did not improve the model fit. Relevance was once again Helmert-coded to test Neutral versus Irrelevant and Relevant versus not-Relevant. Cue was sum-coded. Combined, the contrast coding schemes render the model intercept as the grand mean of Responses in log odds. The cross-level interaction between Cue and Relevance did not significantly improve model fit and, therefore, was not included in the final model. Table 3 presents a regression table for the model of Cue-based Responses for Experiment 2.

The model revealed a significant effect of Cue. There were significantly more Cue-based Responses for ambiguous trials presented with Labels ($M_{\mathrm{L}} = 0.75$) as compared to Numbers ($M_N = 0.64$, $\beta = -0.595$, $SE = 0.260$, $z = 2.291$, $p = .0219$). For the Relevance main

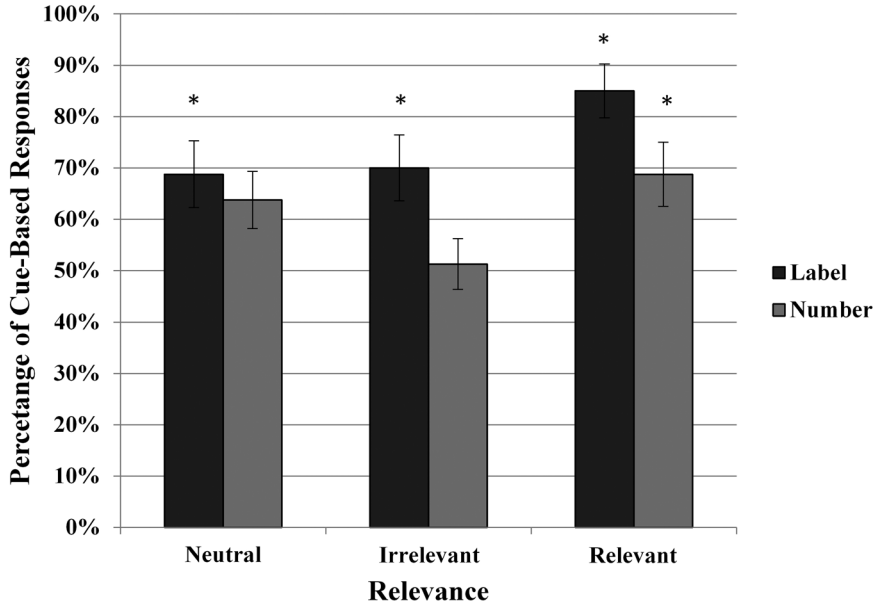

Fig. 4. Percentage of Cue-based Responses for ambiguous trials in Experiment 2. Asterisks reflect significant differences from chance (50%) performance ($p < .05$).

Table 3
Multilevel mixed-effects logistic regression model of Cue-based Responses for Experiment 2

Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	z	p
Intercept	0.9526	0.1400	6.807	1.00e-11***
Cue	0.5952	0.2598	2.291	.021943*
Relevance (Neutral vs. Irrelevant)	0.4700	0.3071	1.531	.125989
Relevance (Relevant vs. not-Relevant)	0.6509	0.2835	2.296	.021659*

Note. *$p < .05$; **$p < .01$; ***$p < .001$.

effect, the model found a significant effect of Relevant versus not-Relevant ($\beta = 0.651$, $SE = 0.284$, $z = 2.296$, $p = .0217$) but did not find a significant difference between the Neutral and Irrelevant conditions ($\beta = 0.470$, $SE = 0.307$, $z = 1.530$, $p = .126$).

As Fig. 4 shows, performance with Labels was significantly different from chance in all conditions (Neutral: $M_L = 0.69$, $t(19) = 2.88$, $p < .05$; Irrelevant: $M_L = 0.70$, $t(19) = 3.11$, $p < .01$; Relevant: $M_L = 0.85$, $t(19) = 6.66$, $p < .001$). Performance with Numbers was at chance in the Irrelevant ($M_N = 0.51$, $t(19) = 0.25$, $p > .05$) and Neutral ($M_N = 0.64$, $t(19) = 2.46$, $p = .05$) conditions but differed from chance in the Relevant ($M_N = 0.69$, $t(19) = 3.00$, $p < .01$) condition.

3.2.2. Memory performance

Table 4 shows the overall accuracy of the memory task. Accuracy was calculated using the target items and within-category distractors only since people rarely (about 2% of the time)

Table 4
Accuracy on memory task of Experiment 2 calculated as the number of correct responses to targets and within-category distractors divided by the total number of targets and within-category distractors

	Label	Number
Neutral	69%	73%
Irrelevant	64%	72%
Relevant	70%	73%

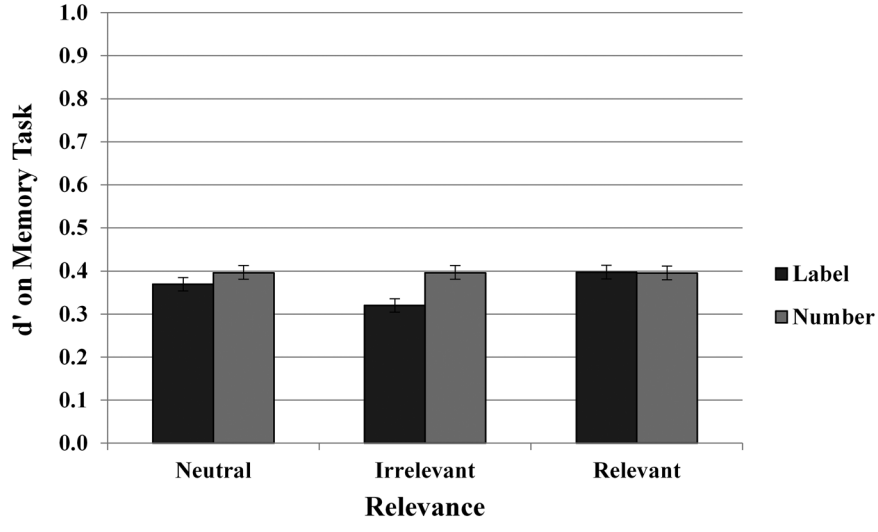

Fig. 5. Memory task performance (d' score) in Experiment 2. Error bars represent SEM.

picked distractors from outside of their assigned Cue category. A d' score was calculated for each participant by subtracting the *z*-score for False Alarm rate on the within-category distractor items from the Hit rate to target items. In order to calculate *z*-scores, Hit and False Alarm rates of 0 and 1 were replaced using the formulas 0.5/N and $(N - 0.5)/N$, respectively, where N was equal to the number of trials used to calculate the Hit or False Alarm rate (Stainslaw & Todorov, 1999). Using the d' score on the memory task as the dependent measure, we performed a two-way ANOVA with Cue and Relevance as factors. Fig. 5 summarizes the data. The ANOVA found no significant main effect of either Cue, $F(1, 114) = 2.378$, $p = .126$ or Relevance, $F(2, 114) = 1.036$, $p = .358$, and no interaction, $F(2, 114) = 1.092$, $p = .339$. These results suggest that there was no difference in participants' ability to recall both cues across conditions. As a result, it is hard to attribute differences between cue types in the categorization task to a lack of encoding of the cues themselves.

3.3. Discussion

Experiment 2 broadly replicated a main result from Experiment 1: Even in a more stringent categorization task, labels had an advantage over other cues (here, numbers) in supporting the

categorization of novel object kinds as predicted by the position that nominal labels mark categories. The advantage of labels persisted even when labels and other potentially helpful cues were equally likely to be remembered in a post-categorization probe. Furthermore, labels were used to constrain categorization choices consistently regardless of their stated task relevance. By contrast, numbers were reliably used only when their relevance to the task was explicitly highlighted (Relevant condition) but not when it was said to be absent (Irrelevant condition) or remained unspecified (Neutral condition).

Two observations about these patterns merit some discussion. First, the fact that, in both Experiments 1 and 2, labels drove categorization at above chance rates even in the Irrelevant condition is unexpected. This might mean that people did not believe that the messages on the screen in the Irrelevant condition were the result of a "glitch," especially for novel labels that could be taken as names for the new objects. If so, a new version of the Irrelevant condition clearly indicating that the cues should be ignored would drive a stronger contrast to the Neutral condition (especially given the lack of a difference between Irrelevant and Neutral conditions in Experiment 2). Second, the surprise memory check in Experiment 2, even though useful, did not provide insight into whether participants were attending to the cues at the time of performing the categorization task. Evidence that people attended to the cues at that moment but used them selectively during categorization would add to the argument that labels have a unique role in guiding kind membership. Experiment 3 was designed to address both of these issues.

4. Experiment 3

In Experiment 3, we created a new version of Experiment 2 that (a) connected instructions in the Relevant and Irrelevant conditions more explicitly to task relevance (and omitted the "glitch" cover story), and (b) added a memory check during the main categorization trials to see whether participants detected and remembered the cues close to the moment of categorization. The new experiment also included slightly amended wording to more clearly orient participants toward the task of categorization (alongside other smaller changes).

4.1. Methods

4.1.1. Participants

A total of 120 English-speaking adults participated in the study online. They were recruited from the subject pool at the University of Pennsylvania and were given partial course credit for participating. A total of 30 originally recruited participants had to be replaced as they either did not accurately detect both cues during the memory checks for the four critical trials ($N = 17$) and/or failed to respond to more than one ambiguous trial ($N = 13$).

4.1.2. Stimuli and procedure

The stimuli and procedure were similar to those of Experiment 2 with the exception of the changes noted below. Participants completed the task individually on PennController for

IBEX (PCIbex), an online platform for conducting experiments (Zehr & Schwarz, 2018). To ensure that participants could properly complete the task in the online environment, we lengthened the time of presentation of the images. Specifically, the triad display in the study phase was presented for a total of 13 s. The images then disappeared for 1 s, but the solid line remained. The images then reappeared with a red border surrounding the display and stayed on the screen during the test phase for 10 s. Participants automatically moved to the next trial after the 10-s test phase. We excluded participants if they did not answer at least three out of the four ambiguous trials within that window. Moreover, we only examined the trials that were answered (there were only 54 unanswered, or "missed," trials across 2400 total trials in the final sample). In a continued effort to ensure that the task was interpreted as involving categorization, we asked participants to "mark which of the top two items the bottom item belongs with" (as opposed to "goes with," Experiment 2) by using their cursor "to choose 'L' if it belongs with the top left item or 'R' if it belongs with the top right item."

Unlike Experiment 2, instructions emphasized task relevance explicitly in both the Irrelevant condition ("As you go through the items, you may see things that look like messages displayed next to some of the images. Please disregard them as they are irrelevant for purposes of your task") and the Relevant condition ("As you go through the items, you may see useful messages displayed next to some of the images. Please take them into account as they are relevant for purposes of your task"). As in the previous experiment, participants in the Neutral condition were not told anything about the messages.

Finally, to ensure that participants across all conditions would attend to all cue types as they performed the task, we added five memory checks during the main experiment. These memory checks occurred after each of the four ambiguous (critical) trials as well as during one additional unambiguous trial. During the memory check, participants were asked to select "yes" or "no" to the question, "Did you see any messages on the screen?" Additionally, they were asked, "If yes, click on the message or messages below" and were shown the full list of eight Label or Number cues that would be used during the experiment. We excluded and replaced participants who did not accurately choose both cues that had been earlier displayed on the screen in each of the four ambiguous trials (see Participants section for such exclusions). Thus, any non-compliance with the cues during categorization could not be attributed to failure to attend to and encode the cues in the first place. This study took about 30 min to complete.

4.2. Results

As in Experiments 1 and 2, we examined the percentage of correct (i.e., perception-driven) responses for the unambiguous trials that lacked cues (10%, 30%, 70%, and 90%). The results were near ceiling for all Cue ($M_L = 0.98$, $M_N = 0.95$) and Relevance ($M_N = 0.97$, $M_I = 0.96$, $M_R = 0.95$) subgroups of participants. For unambiguous trials that were accompanied by cues, participants were near ceiling for all Cue ($M_L = 0.98$, $M_N = 0.95$) and Relevance ($M_N = 0.98$, $M_I = 0.96$, $M_R = 0.94$) manipulations. Thus, participants were able to categorize the unambiguous stimuli with or without additional cues as in Experiments 1 and 2.

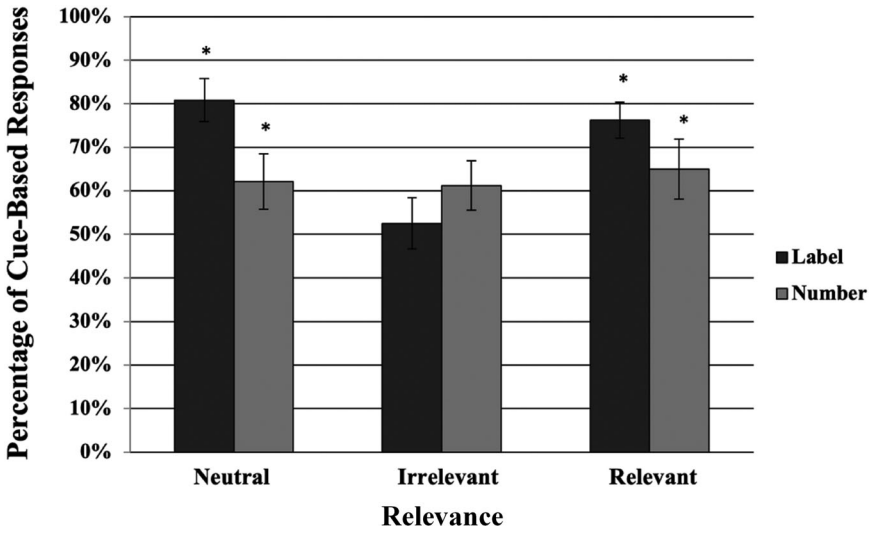

Fig. 6. Percentage of Cue-based Responses for ambiguous trials in Experiment 3. Asterisks reflect significant differences from chance (50%) performance ($p < .05$).

Table 5
Multilevel mixed-effects logistic regression model of Cue-based Responses for Experiment 3

Coefficient	$\hat{\beta}$	$SE(\hat{\beta})$	z	p
Intercept	0.7800	0.1264	6.171	6.77e-10***
Cue	0.4452	0.2419	1.840	.065757
Relevance (Neutral vs. Irrelevant)	0.7842	0.2963	2.647	.008131**
Relevance (Relevant vs. not-Relevant)	0.2964	0.2560	1.158	.246900
Cue:Relevance (Neutral vs. Irrelevant)	1.4320	0.5937	2.412	.015871*
Cue:Relevance (Relevant vs. not-Relevant)	0.2497	0.5116	0.488	.625466

Note. *$p < .05$; **$p < .01$; ***$p < .001$.

For ambiguous (50%) trials, the analysis dataset for Experiment 3 consisted of a total of 471 observations (only nine out of the 480 such trials were missed). Fig. 6 summarizes the data. Our analysis used a mixed-effects logistic regression model that included Cue-based Responses to ambiguous trials as the binary dependent variable, Cue (Label, Number) and Relevance (Neutral, Irrelevant, Relevant) as fixed predictors, and a random intercept for participants. Contrast coding of the predictors followed that of Experiment 2. Table 5 presents a regression table for the model of Cue-based Responses for Experiment 3.

There was no significant main effect of Cue ($\beta = 0.445$, $SE = 0.242$, $z = 1.840$, $p = .066$). There was a main effect of Relevance in the Neutral versus Irrelevant term ($\beta = 0.784$, $SE = 0.296$, $z = 2.647$, $p = .008$), such that participants were overall less reliant on the cue in the Irrelevant condition, compared to the Neutral condition. There was no overall difference between the Relevant condition, compared to the not-Relevant (Neutral and Irrelevant) condition ($\beta = 0.296$, $SE = 0.256$, $z = 1.158$, $p = .247$). Importantly, there was a significant interaction between Cue and whether the Relevance was set to Neutral versus Irrelevant (β

= 1.432, $SE = 0.594$, $z = 2.412$, $p = .016$). To determine the nature of the interaction, we conducted additional paired two-sided t-tests between Label and Number cues within each of the two Relevance conditions. Cue-compliance for Label was higher than for Number in the Neutral condition ($M_L = 0.81$; $M_N = 0.62$; $t(19) = 2.721$, $p < .05$) but the difference disappeared in the Irrelevant condition ($M_L = 0.53$; $M_N = 0.61$; $t(19) = 0.839$, $p = .412$). The interaction term between Cue and the other Relevance term (Relevant vs. not-Relevant) was not significant ($\beta = 0.250$, $SE = 0.512$, $z = 0.488$, $p = .625$).

As Fig. 6 shows, performance with Labels was significantly different from chance in the Neutral and Relevant conditions (Neutral: $M_L = 0.80$, $t(19) = 5.26$, $p < .001$; Relevant: $M_L = 0.76$, $t(19) = 3.68$, $p < .01$) but not the Irrelevant condition ($M_L = 0.53$, $t(19) = 0.41$, $p > .05$). Similarly, performance with Numbers was significantly different from chance in the Neutral and Relevant condition (Neutral: $M_N = 0.62$, $t(19) = 2.30$, $p < .05$; Relevant: $M_N = 0.65$, $t(19) = 3.27$, $p < .01$) but not in the Irrelevant condition ($M_N = 0.61$, $t(19) = 1.69$, $p > .05$).

4.3. Discussion

Experiment 3 extended and refined the broad picture emerging from Experiments 1 and 2. Even though there was no global advantage for labels, compared to numbers, the two cues behaved differently when presented as Neutral versus Irrelevant. As we expected, labels were more likely than numbers to be utilized for categorization when no information was provided about their task relevance but did not differ from numbers when people were explicitly told that they were both irrelevant (presumably because our new Irrelevant condition convinced people to ignore the cues more than prior incarnations in our two previous experiments). Labels and numbers did not have differential effects depending on whether cues were presented as Relevant versus not-Relevant. Since in-task memory tests ensured that people attended to both cues across conditions, the fact that labels and numbers had different effects on aspects of categorization depending on task relevance points to underlying semantic factors that link nominals (but not numerals) to object kind membership.

As in Experiments 1 and 2, participants used the novel labels during categorization at levels different from chance in both the Relevant and Neutral conditions. However, given our modified instructions, participants no longer did so in the Irrelevant condition. Unlike in the previous experiments, participants used numbers reliably in the Relevant *and* Neutral (but not the Irrelevant) conditions. Thus, in this context that lacked more specific task-relevant information, both label and non-label cues were reliably treated as potentially relevant pieces of information (cf. Sperber & Wilson, 1986), especially for solving an otherwise ambiguous task. Beyond this similarity, however, as mentioned already, labels were used more frequently than numbers in drawing category boundaries for novel natural kind exemplars in such underspecified (Neutral) contexts.

5. General discussion

Several studies in the literature have demonstrated that adults use linguistic labels to draw the boundaries of newly encountered categories (e.g., Deng & Sloutsky, 2012; Johanson

& Papafragou, 2016; Lupyan et al., 2007; Lupyan & Thompson-Schill, 2012; Yamauchi & Markman, 2000). In most, if not all, prior demonstrations of the role of labels on adult categorization, the contribution of the representational content of the labels cannot be easily distinguished from the experimenter's intention to direct attention to the labels (and presumably their task relevance) within the categorization task. For the same reason, the potency of labels in adults remains to be compared to that of other cues that are also introduced as relevant.

Clarifying the source and boundaries of the effects of labels on category formation is theoretically important, especially for the widely held perspective that labels are inherently category markers and have a privileged connection to kinds because of their representational, or semantic, content (Balaban & Waxman, 1997; Gelman & Davidson, 2013; Gelman & Markman, 1986; among others). This position can naturally accommodate the fact that the use of labels during categorization should depend on the conditions in which the labels are presented and that other cues might also shape the categorization of novel kinds if accompanied by strong invitations to use these cues within the task. This position expects that labels (because of their representational nature) should shape category formation even if the experimenter does not specifically direct participants' attention to the labels and/or their potential significance within the task (and leaves open the possibility that other cues might do so if participants infer that the cues are meant to be relevant; Sperber & Wilson, 1986; see Ferguson & Waxman, 2016). Crucially, this view predicts that, even when relevance is held constant, adults should assume that labels are uniquely helpful for categorization, compared to other symbolic cues that do not have a privileged representational connection to kinds, other things being equal. The present study was the first to compare the role of different levels of relevance and different cue types (Experiment 1.: labels, numbers, symbols; Experiments 2 and 3: labels, numbers) in adult categorization with the goal of testing these predictions. To isolate the effects of labels and other cues more clearly, we asked adults to make category decisions when perceptual information provided no category boundaries for novel exemplars of natural kinds (i.e., in ambiguous trials).

These predictions of the position that labels are category markers were confirmed in our data. Experiment 1 showed that several types of cues (i.e., novel labels, numbers, and symbols) were used to aid the formation of a novel natural kind category when they were assumed to be relevant to the task. However, labels had an overall advantage in shaping the categorization of novel natural kinds, compared to numbers or arbitrary symbols, regardless of whether people were explicitly invited to use the cues or not. Experiment 2 showed that the label advantage over numbers persisted in more stringent categorization tasks; furthermore, that advantage could not be explained by assuming that labels were simply more salient than numbers. Experiment 3 found that, when the task (ir)relevance of cues was clearly introduced, there was still room for a bias favoring labels over numbers when no specific instructions were given about the usefulness of the cues; furthermore, this bias could not be explained by higher attention to labels than numbers during the categorization task. Overall, labels were preferentially used by adults, compared to other symbolic stimuli, to categorize natural kinds even when perceptual information provided no clear category boundaries.

Taken together, our findings show that the effect of labels during categorization cannot simply be attributed to broad effects of communication that could be achieved with any symbolic representation. Even though communication in general carries a presumption of relevance and is therefore expected to yield "cognitive effects" for the hearer (Sperber & Wilson, 1986), linguistic (noun) labels create the expectation of a specific type of cognitive effect—here, kind reference—that is not readily shared by other meaningful cues (e.g., numbers). By the same token, our findings show that adults do not approach the categorization task through a strategy to use whatever cue they are presented with to base their category decisions on (Goldstone et al., 2001). For similar reasons, our results cannot be explained solely by the idea that labels function by directing people's attention toward certain kinds of groupings of exemplars over others and encouraging the extraction of similarities (e.g., Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002) since in principle all cues in our studies could function in that way. The specific and unique effects of labels suggest that adults use sophisticated and rich reasoning to determine which kinds of information bear on natural kind membership and under what circumstances—exactly as predicted by the position that labels are privileged types of invitations to form categories.

Recall that a newly presented label in Experiments 1 and 2 was used to guide the categorization of unfamiliar stimuli at above-chance levels even when people were given reason to treat the label as accidental (Irrelevant condition). This result might appear surprising, even unexpected, under the theoretical position that humans treat labels as category indicators. This finding might also appear at odds with developmental work suggesting that, when the intentional link between a label and a category is severed (e.g., when the source of the label is non-human), even very young learners recognize that labels should not be linked to categories (Campbell & Namy, 2003; Fulkerson & Haaf, 2003; see also Gelman & Bloom, 2000; Jaswal, 2004; Tomasello & Barton, 1994; Tomasello, Strosberg, & Akhtar, 1996). We suspect that, given the perceptually ambiguous nature of our natural kind stimuli, participants did not completely believe the "glitch" story in that condition and instead relied on labels to resolve the category ambiguity of the stimuli. Recall that these stimuli could still be taken as the referents of these labels (since there was no mismatch between the labels and the visual evidence for the structure of the novel categories). Interestingly, labels were not used to guide categorization in a redesigned Irrelevant condition in Experiment 3 where the instructions clearly pointed out that the cues were not helpful and should be ignored (and omitted the "glitch" cover story).

The present studies included manipulations of task relevance that were purposefully removed from actual interactions with an interlocutor. Nevertheless, throughout all experiments, when global information about the experimenter's intent established the relevance of symbolic cues (numbers or arbitrary symbols), such cues guided kind membership at rates different from chance (Relevant condition). Furthermore, in Experiment 3, even in the absence of explicit information, symbolic cues such as numbers were used as a signal to inform how new instances of natural kinds should be grouped into categories (Neutral condition). This suggests that in certain contexts, non-label cues may also guide category membership. After all, numbers at an airport can indicate terminals or boarding groups, which have referential power. Nevertheless, these uses of numbers require specific conventions to be successful.

The methodological paradigm leveraged here has been used extensively in the literature to investigate various aspects of human categorization. However, we propose that to properly interpret the results, it is necessary to integrate thinking about how the cues are actually interpreted in communication. Our findings show that participants are sensitive to the presumed relevance of a variety of ostensive stimuli (Sperber & Wilson, 1986) and align with other evidence that, in addition to language, a host of non-linguistic communicative cues can help establish novel object categories (Perszyk & Waxman, 2018). Furthermore, this pattern coheres with the strong role of goal directedness, especially in communicative conditions, for learning novel information in both adults (Stephens, Perfors, & Navarro, 2010) and infants (Gergely & Csibra, 2013). However, even within strongly relevant or neutral contexts, reliance on labels surpassed reliance on other cues when inferring novel category structure (see especially Experiment 3).

Naturally, the present data have limitations. We have studied categorization with undergraduate, English-speaking, and college-aged students and our findings would have to be extended to additional populations (as well as young children, see Introduction). Furthermore, even though this type of paradigm has been massively used to study reliance on categories in human thinking, it needs to be extended to cases where the processes of categorization serve everyday purposes of thinking about the world.

Acknowledgments

This work was partly supported by NSF Grant #1632849. We thank Hannah Schwarz, Ariel Mathis, Alessandra Pintado-Urbanc, and especially June Choe for their help in collecting and analyzing the data. The first author is now at the U.S. Army Research Laboratory, and the second author is now at Mayo Clinic.

Notes

1 To ensure that FantaMorph ratings agreed with human intuitions, a separate group of 18 undergraduate students rated the perceptual similarity of all targets to the standards (cf. also Johanson & Papafragou, 2016). Triads were created for each morphed object with the two Standards on top and a Target below them. Participants viewed all possible triads in one of two randomized orders and were asked to mark where the target fell on a 9-point scale from 10% to 90%, with Standard 1 at 0% and Standard 2 at 100%. The ratings were averaged across adults for targets at each 10% interval (as determined by FantaMorph). The only cases where the mean ratings were different from the expected ratings were the 10% target ($M = 11.40$, $t(17) = 3.37$, $p = .004$), and the 90% target ($M = 84.03$, $t(17) = -3.98$, $p = .001$), but even so these objects were rated as unambiguously closer to the appropriate standard. Overall, the FantaMorph ratings capture human intuitions about perceptual similarity.

2 In the data reporting, we use two sets of subscripts. For Cue: Label (L), Number (N), and Symbol (S), and for Relevance: Neutral (N), Irrelevant (I), and Relevant (R).

References

Althaus, N., & Plunkett, K. (2015). Timing matters: The impact of label synchrony on infant categorisation. *Cognition*, *139*, 1–9.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429.

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, *64*(1), 3–26.

Booth, A. E., & Waxman, S. R. (2002a). Object names and object functions serve as cues to categories for infants. *Developmental Psychology*, *38*(6), 948–957.

Booth, A. E., & Waxman, S. R. (2002b). Word learning is 'smart': Evidence that conceptual information affects preschoolers' extension of novel words. *Cognition*, *84*(1), B11–B22.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108*, 10918–10925.

Campbell, A. L., & Namy, L. L. (2003). The role of social-referential context in verbal and nonverbal symbol learning. *Child Development*, *74*(2), 549–563.

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

Casasola, M., & Bhagwat, J. (2007). Do novel words facilitate 18-month-olds' spatial categorization? *Child Development*, *78*, 1818–1829.

Deng, W., & Sloutsky, V. M. (2012). Carrot eaters or moving heads: Inductive inference is better supported by salient features than by category labels. *Psychological Science*, *23*, 178–186.

Fairchild, S., Mathis, A., & Papafragou, A. (2018). Linguistic cues are privileged over non-linguistic cues in young children's categorization. *Cognitive Development*, *48*, 167–175.

Fairchild, S., & Papafragou, A. (2019). Language and categorization in monolinguals and bilinguals. *Bilingualism: Language and Cognition*, *23*(3), 618–630.

Ferguson, B., & Waxman, S. (2016). Linking language and categorization in infancy. *Journal of Child Language*, *44*(3), 527–552.

Fisher, A. V. (2010). What's in the name? Or how rocks and stones are different from dogs and puppies. *Journal of Experimental Child Psychology*, *105*, 198–212.

Fulkerson, A. L., & Haaf, R. A. (2003). The influence of labels, non-labeling sounds, and source of auditory input on 9- and 15-month-olds' object categorization. *Infancy*, *4*, 349–369.

Fulkerson, A., & Waxman, S. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition*, *105*(1), 218–228.

Gelman, S. A., & Bloom, P. (2000). Young children are sensitive to how an object was created when deciding what to name it. *Cognition*, *76*(2), 91–103.

Gelman, S. A., & Davidson, N. S. (2013). Conceptual influences on category-based induction. *Cognitive Psychology*, *66*, 327–353.

Gelman, S. A., & Markman, E. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Gergely, G., & Csibra, G. (2013). Natural pedagogy. In M. R. Banaji & S. A. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 127–132). Oxford, England: Oxford University Press.

Gliozzi, V., Mayor, J., Hu, J.-F., & Plunkett, K. (2009). Labels as features (not names) for infant categorization: A neurocomputational approach. *Cognitive Science*, *33*, 709–738.

Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*, 178–200.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27–43.

Hoffman, A. B., & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General*, *139*, 319–340.

Jaswal, V. K. (2004). Don't believe everything you hear: Preschoolers' sensitivity to speaker intent in category induction. *Child Development*, *75*, 1871–1885.

Jaswal, V. K., & Markman, E. M. (2007). Looks aren't everything: 24-month-olds' willingness to accept unexpected labels. *Journal of Cognition and Development*, *8*, 93–111.

Johanson, M., & Papafragou, A. (2016). The influence of labels and facts in children's and adults' categorization. *Journal of Experimental Child Psychology*, *144*, 130–151.

Landau, B., & Shipley, E. (2001). Labeling patterns and object naming. *Developmental Science*, *4*, 109–118.

Lupyan, G., Rakison, D., & McClelland, J. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*, 1077–1083.

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

Lupyan, G., & Spivey, M. (2008). Perceptual processing is facilitated by ascribing meaning to novel stimuli. *Current Biology*, *18*(10), R410–R412.

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

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592–613.

Perfors, A., & Navarro, D.J. (2010). How does the presence of a label affect attention to other features? In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 1834–1839). Austin, TX: Cognitive Science Society.

Perszyk, D.R., & Waxman, S.R., (2018). Linking language and cognition in infancy. *Annual Review of Psychology*, *69*, 231–250.

Plunkett, K., Hu, J.-F., & Cohen, L. (2008). Labels can override perceptual categories in early infancy. *Cognition*, *106*, 665–681.

Sloutsky, V. M., Lo, Y., & Fisher, A. V. (2001). How Much Does a Shared Name Make Things Similar? Linguistic Labels, Similarity, and the Development of Inductive Inference. *Child Development*, *72*(6), 1695–1709.

Smith, L.B., Jones, S.S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*(1), 13–19.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and Cognition*. Oxford, England: Blackwell.

Stainslaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior, Research Methods, Instruments, & Computers*, *31*(1), 137–149.

Stephens, R. G., Perfors, A., & Navarro, D. J. (2010). Social condition effects on the impact of category labels. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual meeting of the cognitive science society* (pp. 1411–1416). Austin, TX: Cognitive Science Society.

Tomasello, M., & Barton, M. (1994). Learning words in nonostensive contexts. *Developmental Psychology*, *30*, 639–650.

Tomasello, M., Strosberg, R., & Akhtar, N. (1996). Eighteen-month-old children learn words in non-ostensive contexts. *Journal of Child Language*, *22*, 1–20.

Ünal, E., & Papafragou, A. (2016). Interactions between language and mental representations. *Language Learning*, *66*(3), 554–580.

Waxman, S. R., & Markow, D. B. (1995). Words as invitations to form categories: Evidence from 12-month-old infants. *Cognitive Psychology*, *29*, 257–302.

Welder, A.N., & Graham, S.A. (2006). Infants' categorization of novel objects with more or less obvious features. *Cognitive Psychology*, *52*, 57–91.

Yamauchi, T., Kohn, N., & Yu, N.-Y. (2007). Tracking mouse movement in feature inference: Category labels are different from feature labels. *Memory & Cognition*, *35*, 852–863.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, *39*, 124–148.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 776–795.

Yamauchi, T., & Yu, N.-Y. (2008). Category labels versus feature labels: Category labels polarize inferential predictions. *Memory & Cognition*, *36*, 544–553.

Zehr, J., & Schwarz, F. (2018). *PennController for Internet Based Experiments (IBEX)*. Retrieved February 20, 2023, from: https://doi-org.proxy.library.upenn.edu/10.17605/OSF.IO/MD832

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Experiment 1 – Categorization task data.
Experiment 2 – Categorization task data.
Experiment 2 – Memory task data.
Experiment 3 – Categorization task data.