# Continuity and change in the development of category-based induction: The test case of diversity-based reasoning

Marjorie Rhodes [a,*], Peter Liebenson [b]

[a] New York University, 6 Washington Place, 301, New York, NY 10003, United States
[b] New York University, 6 Washington Place, 302, New York, NY 10003, United States

### ARTICLE INFO

### ABSTRACT

The present research examined the extent to which the cognitive mechanisms available to support inductive inference stay constant across development or undergo fundamental change. Four studies tested how children (ages 5–10) incorporate information about sample composition into their category-based generalizations. Children's use of sample composition varied across age and type of category. For familiar natural kinds, children ages 5–8 generalized similarly from diverse and non-diverse samples of evidence, whereas older children generalized more broadly from more diverse sets. In contrast, for novel categories, children of each age made broader generalizations from diverse than non-diverse samples. These studies provide the first clear evidence that young children are able to incorporate sample diversity into their inductive reasoning. These findings suggest developmental continuity in the cognitive mechanisms available for inductive inference, but developmental changes in the role that prior knowledge plays in shaping these processes.

© 2015 Elsevier Inc. All rights reserved.

## 1. Introduction

Category-based induction is a fundamental process by which humans acquire and extend knowledge (Murphy, 2002; Rips, 1975). By allowing people to overlook superficial differences and focus

---

* Corresponding author.
   *E-mail addresses:* marjorie.rhodes@nyu.edu (M. Rhodes), peter.liebenson@nyu.edu (P. Liebenson).

on underlying regularities, categories enable efficient learning (e.g., allowing people to assume that because one dog likes to be petted, others will too, despite differences in size or color). Examining the development of category-based induction thus has the potential to answer central questions in cognitive development—particularly whether the mechanisms that underlie knowledge acquisition are continuous across development or undergo fundamental change (Hayes, 2007).

Efficient category-based induction involves determining a reasonable scope of generalization from a particular set of evidence. Normative models of induction indicate that a key feature that contributes to this determination is *sample diversity* (Heit, 2000; Kemp & Tenenbaum, 2009)—diverse samples drawn from a category (e.g., a tiger and a mouse) warrant strong generalizations to broader kinds (e.g., mammals), whereas less diverse samples (e.g., a tiger and a lion) warrant generalization to more limited subsets (e.g., felines). Diverse samples support broader generalizations because they provide more coverage of the category of interest (Osherson, Smith, Wilkie, López, & Shafir, 1990). Consideration of sample diversity is consistent with a key principle of the philosophy of science—that before drawing general conclusions, it is useful to obtain evidence from diverse sources (Bacon, 1620/1898; Nagel, 1939; see Heit, Hahn, & Feeney, 2005).

Consistent with this principle, adults make stronger generalizations based on diverse samples of evidence (a phenomenon referred to as "diversity-based reasoning"). For example, adults rate inferences based on diverse samples of animals (e.g., "dogs and whales have four-chambered hearts, therefore all mammals have four-chambered hearts") as stronger than those based on less diverse samples of animals (e.g., "dogs and wolves have four-chambered hearts, therefore all mammals have them"; Osherson et al., 1990; also Feeney & Heit, 2011). Adults also seek out diverse evidence to determine if there is support for category-wide generalizations, preferring to assemble samples composed of a lion and a hamster, rather than a lion and a tiger, for example, to discover properties of mammals (Kim & Keil, 2003; López, 1995; López, Atran, Coley, Medin, & Smith, 1997; Rhodes, Brickman, & Gelman, 2008; Rhodes, Gelman, & Brickman, 2008).

Sample diversity is not the only criterion that adults use to evaluate evidence—they can also rely on other sample features (e.g., typicality, sample size; Osherson et al., 1990), as well as other types of knowledge, especially knowledge of the causal mechanisms involving particular properties (Medin, Coley, Storms, & Hayes, 2003; Proffitt, Coley, & Medin, 2000). Yet, sample diversity is an important factor that adults consider across many experimental tasks (Heit, 2000; Heit et al., 2005). Although there is variation across tasks and participants in *when* adults appeal to diversity to evaluate samples (e.g., experts vs. non-experts; participants from different cultural contexts; Medin et al., 2003), the *ability* to incorporate sample diversity into inductive reasoning has been found in every adult population studied to date (Atran, 1990; Feeney & Heit, 2007; López et al., 1997). Thus, examining how children incorporate sample diversity into their inductive reasoning across development provides a useful test case of whether the mechanisms that support induction stay constant across development or undergo fundamental change.

Here we examine the consideration of sample diversity in inductive reasoning across development, with the aim of identifying whether performance differences across childhood (as reported in prior work) reflect fundamental changes in the mechanisms available to support inductive inference. In contrast to adults, children below the age of 9 often generalize equivalently from diverse and non-diverse samples in experimental tasks. For example, children between the ages of 5 and 8 who learned that three diverse birds have hollow bones (e.g., an eagle, owl, and robin) were no more likely to infer that "all birds have hollow bones" than those who learned that three non-diverse birds had them (e.g., three robins; Gutheil & Gelman, 1997; Li, Cao, Li, Li, & Deák, 2009; López, Gelman, Gutheil, & Smith, 1992). Younger children also do not consider sample diversity when selecting evidence; for example, 6-year-olds are just as likely to select a non-diverse sample (e.g., two Dalmatians) as a diverse sample (e.g., a Dalmatian and a pit bull) to find out if something is true of a category (e.g., dogs; Rhodes, Brickman et al., 2008; Rhodes, Gelman, & Brickman, 2010; Rhodes, Gelman et al., 2008; Zhong, Lee, Huang, & Mo, 2014). In these studies, diversity-based reasoning has been found to emerge around age 8 and become more robust by ages 10–11.

Although there have been a few positive reports of diversity effects in younger children (Heit & Hahn, 2001; Shipley & Shepperson, 2006; Zhong et al., 2014), careful consideration of the experimental tasks used in those reports indicates that children may have shown preferences for diverse samples without actually viewing diverse evidence as more informative. For example, Heit and Hahn (2001)

presented 5-year-olds with a set of diverse dolls belonging to one character and a set of non-diverse dolls belonging to another. Children were presented with another different doll, and asked to guess who owned it. Children reliably inferred that the person owning the diverse dolls owned the new doll. Although this is an interesting finding, children could solve this task by recognizing that diverse items better match diverse than non-diverse sets, rather than by recognizing that diverse samples support stronger inferences or broader scopes of generalization (for further discussion of this work and other reports of positive diversity effects in children, see Rhodes, 2012; Rhodes, Brickman, et al., 2008; Rhodes, Gelman, et al., 2008).

Thus, to date, there is no clear evidence that children below the age of eight recognize the informative value of diverse evidence, perhaps indicating that the mechanisms underlying inductive reasoning change fundamentally across development. For example, younger children might have difficulty generating the relevant inclusive categories from particular samples (López et al., 1992), computing the relation between samples and abstract categories (Gutheil & Gelman, 1997; Li et al., 2009), or considering how the informative value of samples containing multiple pieces of evidence differs from the informative value of each piece of evidence considered separately (Rhodes, Brickman, et al., 2008). From this perspective, younger children simply lack some cognitive skill that is necessary to incorporate diversity into their inductive reasoning.

Several pieces of recent evidence challenge this explanation, however. First, young infants can make sophisticated inferences about the relation of samples to populations; for example, in the first year of life, infants infer that a sample composed of a particular ratio of red and white balls likely came from a distribution that reflects that composition (Denison, Reed, & Xu, 2013; Xu & Garcia, 2008). These findings cast doubt on the proposal that older children lack the ability to reason about how the composition of a sample reflects a broader population. Further, Rhodes et al. (2010) found that 5-year-olds considered sample diversity to draw an inference about a communicator's intent (see also Xu & Tenenbaum, 2007), but not—in a very similar paradigm—to draw inferences about the state of the world. In this study, 5-year-olds were just as likely to infer that a new property applied to all dogs after learning that it was held by three Dalmatians or three diverse dogs, for example (as in prior work, Guthiel & Gelman, 1997; López et al., 1992). Yet, when the task involved inferring what a teacher meant to teach (e.g., whether the teacher was trying to teach about all dogs or only Dalmatians), children used sample diversity in an adult-like manner (inferring that the property applied only to Dalmatians after seeing the non-diverse sample). As both of these tasks presented identical evidence and involve computing the extent to which a sample represents a broader category, it is difficult to see how deficits in children's basic cognitive abilities could explain their failure in this paradigm.

If young children have the capacities to track and compare sample diversity (Heit & Hahn, 2001; Rhodes, Brickman, et al., 2008; Rhodes et al., 2010) and to reason about how sample composition reflects a broader population (Shipley & Shepperson, 2006; Xu & Garcia, 2008; Xu & Tenenbaum, 2007), why would they consistently fail to incorporate diversity information into their category-based inferences? Answering this question is key to determining the nature of developmental changes in inductive reasoning across childhood. Here we test a new explanation for this developmental change—we propose that young children are capable of engaging in diversity-based reasoning, but that their category knowledge often interferes with these abilities.

Much of the previous research on the development of induction has focused on familiar biological kinds. Biological kinds support many inferences about known and yet-to-be discovered properties, have a clear hierarchical structure, and play an important role in people's daily lives (Carey, 1985; Gelman, 2003; Heit, 2000). Although these qualities make understanding the development of induction involving biological kinds particularly important, we suggest that children's knowledge of these categories—particularly their knowledge of category-typical examples and properties—leads them away from diversity-based reasoning.

Typicality plays a strong role in young children's concepts. Younger children (ages 4–7) learn to classify typical category members before atypical ones (Anglin, 1986; Mervis & Pani, 1980) and generate primarily typical exemplars when asked to name members of a category (Rosner & Hayes, 1977). Further, there is some evidence that young children (age 6) make categorization decisions solely by comparing to abstract prototypes, whereas older children and adults incorporate both prototypes and information about specific, variable exemplars into these decisions (Hayes & Taplin, 1993). Thus

we propose that when young children are asked to consider whether a sample is informative regarding a familiar natural kind, they bring to mind primarily typical exemplars (e.g., robins and blue jays) and seek out samples that share features with the exemplars they have brought to mind (e.g., birds that are relatively small and fly), instead of those that provide coverage of the broader kind. If so, younger children would treat diverse and non-diverse samples as equivalently informative for category-wide generalizations if typicality were held constant across the samples, as has usually been done in previous developmental work.

From this perspective, children fail to engage in diversity-based reasoning for familiar natural kinds not because they lack the cognitive mechanisms necessary to make use of diversity information, but because they view an alternate feature of sample composition—the extent to which a sample contains typical exemplars with highly typical properties—as more informative. In the absence of prior knowledge regarding category-typical features, then, children might instead adopt a coverage-based strategy and utilize diversity-based reasoning. We test this possibility in the present set of studies. Study 1 examines the development of diversity-based reasoning regarding a familiar animal category, to replicate prior work with a new method designed for the present research. Study 2 examines children's reasoning about novel categories. Study 3 compares children's reasoning about familiar and novel categories using identical perceptual stimuli. Study 4 directly tests the relation between typicality beliefs and diversity-based reasoning. These studies include children ages 5–10, spanning the ages at which diversity-based reasoning has been found to develop (Gutheil & Gelman, 1997; López et al., 1992; Rhodes, Brickman, et al., 2008; Rhodes, Gelman, et al., 2008). If children's difficulties with diversity-based reasoning, as reported in prior work, reflect fundamental developmental changes in the mechanisms available to support inductive reasoning, then we should find similar patterns across all studies. Alternately, if category knowledge interferes with diversity-based reasoning in early childhood, then performance—particularly of younger children—should vary by category-type.

## 2. Study 1

Study 1 aimed to replicate prior work showing a slow development across childhood in the emergence of diversity-based reasoning for familiar animal categories. One concern regarding the methods used in previous work is that the induction tasks required children to compare two samples of evidence at once, which might have been both taxing and confusing for younger children (Lawson & Fisher, 2011). For example, in Gutheil and Gelman (1997), children were told that one property was found in a diverse set of birds and another was found in a non-diverse set of birds. Then they were asked to choose which property they would generalize to "all birds." This task is demanding, as children have to track how well each sample represents the broader category and compare these relative strengths. The task could also be confusing, as presenting different properties in the diverse and non-diverse samples implicitly communicates that each property was found in one sample and *not* the other, thus complicating the issue of whether *either* property should be generalized to all members of the kind. These two concerns suggest that young children might perform better if they are able to evaluate samples one at a time. Thus, here children determined an appropriate scope of generalization from diverse or non-diverse samples on separate trials.

### 2.1. Methods

#### 2.1.1. Participants

Participants included 71 children (38 male, 33 female; *M* age = 7.9 years, range 4.7–10.9 years) recruited from and tested at the Discovery Room in the American Museum of Natural History (AMNH). Four additional children began testing but were excluded from analyses: three because they chose not to complete the study and one because of experimenter error.

#### 2.1.2. Training

Children were trained to use a 5-point scale with different numbers of dots to depict "all" (scored as 4), "most" (3), "some" (2), "a few" (1), and "just three" (0; see Fig. 1). To account for any biases that
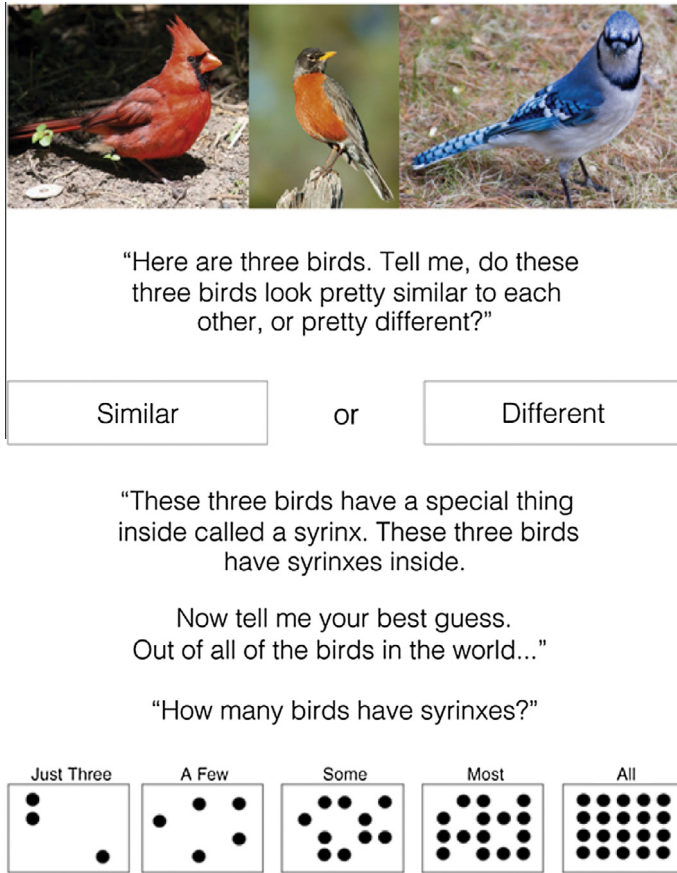
**Fig. 1.** Sample trial, Study 1.

children might have to point more often to the left or right side of the scale, half of the children used a scale that increased from left to right, while half used a scale that increased from right to left. Children answered a series of practice items to become familiar with this scale (e.g., "How many kids have birthdays? Yes, that's right, all kids have birthdays! Point to this card that's all filled up with dots, because that means that all kids have birthdays.") Children completed five practice items, one corresponding to each point on the dot scale. Corrective feedback was provided throughout training as necessary. Full scripts and all stimuli for these experiments are available in the online supplementary materials.

*2.1.3. Induction task*

To begin the induction task, children were told, "We have a lot of different birds here. I'm going to tell you some special things about some of these birds, and your job is to guess how many other birds have those things too. Some of them will be things that all birds have, some of them will be things that only a few birds have, some of them will be things that most birds have, and your job is to take your best guess." To help children keep the category in mind, they were shown a collage containing many birds, and told, "Look at all of these birds!" This sheet remained on the table throughout testing (see Fig. 2).

For each trial (see Fig. 1), children viewed a small sample of birds and heard a new property about the birds in the sample. Then, they were asked to guess how broadly the property might generalize.

**Fig. 2.** Collages to illustrate target categories; Study 1 and bird condition of Study 4, left; Study 2 and "modie" condition of Study 4, center; Study 3, right.

**Table 1**
Properties of the samples in the induction task, by study.

| Studies 1, 3, and 4 | Study 2 |
|---|---|
| have a special thing inside called a syrinx | shrink when it's cold outside |
| see special colors that people don't see | make a rattling noise |
| have something special inside called an air sac | are very warm to the touch |
| need a lot of a special chemical called calcium | can float on water |
| have a special thing inside called a gizzard[a] | make a squeaky noise when it rains |
| have little scales on them | smell like grapefruits |
| have a four-chambered heart | have a little tube on the back called a blicket |
| have hollow bones | light up when it gets dark |
| don't have any teeth | have a big thing inside called a wug |

[a] For studies 3 and 4, this was replaced with, "have a special thing inside called a furcula."

Children completed three items in which the samples contained non-diverse birds (e.g., three robins) and three in which the samples contained diverse birds (e.g., a robin, bluebird, and cardinal; samples were equated for typicality based on ratings obtained from both 5-year-old children and adults in Rhodes, Brickman, et al. (2008) and Rhodes & Gelman, 2009). Items were blocked by sample-type and blocks were presented in a counterbalanced order across participants. To both draw children's attention to the sample composition (to maximize chances for children's success) and document whether they perceived the diverse sample as relatively diverse, each sample was introduced as follows: "Here are three birds. Tell me, do these birds look pretty similar to each other or pretty different?" (scored 1 = pretty different, 0 = pretty similar). No feedback was provided.

Next, the target property was introduced and the induction question was asked (e.g., "These three birds have a special thing inside called a syrinx. These three birds have a syrinx inside. Now tell me your best guess, of all the birds in the world, how many birds have syrinxes?") Children responded verbally or by pointing to the picture on the dot scale that indicated their response. The particular properties rotated across sample-types across children, so that equal numbers of children were asked about "syrinxes" in the diverse and non-diverse samples, for example. For a list of all properties, see Table 1.

On one of the diverse trials and one of the non-diverse trials, children were asked a follow-up question regarding whether they would also generalize the target property to members of a different animal category—frogs. Whether the frog question was asked after the first, second, or third question within each block was counter-balanced across participants. For example, following the sample item described above, children would be asked: "What about frogs? Do you think any frogs have a special thing inside called a syrinx?" If children responded affirmatively, they were asked "How many?" and gave a response using the dot scale. To allow us to compare these responses to those for birds, responses in which children did not generalize or responded with the lowest scale point were both scored as 0, with the remainder of the scale scored as for the bird items. These items allow us to assess whether children of each age (regardless of consideration of sample diversity) engaged in systematic category-based induction (e.g., generalizing more to birds than to more distantly related categories).

For exploratory purposes, children also completed three items in which they were introduced to samples containing a single bird. From a normative perspective, it is not clear how broadly these samples should support generalization relative to the diverse and non-diverse sets. From one perspective, they provide the weakest support for generalizations, as the other two sets contain greater sample sizes (Lawson, 2014; Osherson et al., 1990). From another perspective, they might support more generalizations than the non-diverse sets, as the non-diverse sets could be viewed as providing evidence that a property only applies to a limited subset (i.e., if participants assume strong sampling, Xu & Tenenbaum, 2007). Due to this ambiguity and because our key research question involves how children respond to diverse and non-diverse samples of equivalent sample size, we did not analyze these items with the other sample-types. Children's generalizations from these single samples can be found in Table A1.

### 2.1.4. Control questions

To document that children of each age could properly use the scale, children were asked two questions at the end of the study; one for which they should respond towards the low end of the scale, "Here are three birds. I met these three birds yesterday. I only met these three birds yesterday. Now tell me your best guess, of all the birds in the world, how many birds did I meet yesterday?" and one for which they should use a higher scale point, "Here is a bird. This bird is a girl. Now tell me your best guess, of all the birds in the world, how many birds are girls?"

### 2.1.5. Evidence selection

In order to compare the present findings to prior work, after the induction items, children completed four evidence selection tasks, following the method of Rhodes, Brickman, et al. (2008). Children were told, "Now we are going to play a new game. We're going to pretend you're a scientist and you're going to help me learn new things about animals. Now you are a scientist studying cats. Your job is to find out if cats have papillae. But you can't look at all of the cats in the world to find out about cats. You can only look at two cats, just two. Which cats do you want to look at to find out about cats? Do you want to look at these two cats? Or these two cats?" Children were offered a choice between two non-diverse cats (scored 0) or two diverse cats (scored 1). The side of the diverse and non-diverse samples was balanced across items. Children completed four items involving cats, dogs, fish, and pigs.

## 2.2. Results

### 2.2.1. Generalizations to birds

Children's generalizations to birds from diverse and non-diverse samples were analyzed via repeated measures analysis of variance, with sample (diverse, non-diverse) as a within-subjects variable. Because we had a large and continuous age range, and given mixed reports of age effects in prior work (e.g., Gutheil & Gelman, 1997; Heit & Hahn, 2001; López et al., 1992; Rhodes, Brickman, et al., 2008; Rhodes, Gelman, et al., 2008), we began by considering age as a continuous predictor in the analysis, to test whether further analyses to clarify age-related changes were warranted for this task. This analysis yielded a main effect of sample (diverse vs. non-diverse), $F(1,69) = 5.65$, $p = .02$, and a sample X age interaction, $F(1,69) = 6.93$, $p = .01$. Collapsing across age, children generalized slightly more from diverse ($M = 2.35$, $SE = .11$) than non-diverse ($M = 2.24$, $SE = .12$) samples, Cohen's $D = .11$.

Given that performance varied by age, we next divided children into age groups based on prior work suggesting that diversity-based reasoning emerges around age 8 on some tasks and becomes more robust by age 10 (López et al., 1992; Rhodes, Brickman, et al., 2008; Rhodes, Gelman, et al., 2008). Thus, we divided children into three age groups: ages 5–6 ($n = 26$; $M = 6.0$); ages 7–8 ($n = 23$; $M = 7.9$); ages 9–10 ($n = 22$; $M = 10.0$) and tested for an effect of sample separately within each of these age-groups using paired sample $t$-tests. Among children aged 5–6 and aged 7–8, there were no effects of sample; children aged 9–10, however, generalized more from diverse than non-diverse samples, $t(21) = 2.20$, $p = .04$, Cohen's $D = .49$, see Fig. 3.
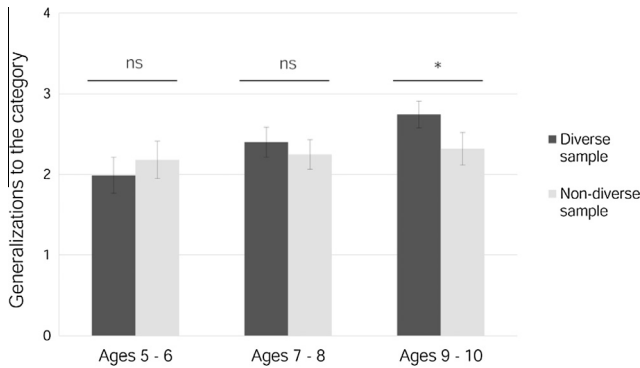
**Fig. 3.** Children's generalizations to birds from diverse and non-diverse samples, by age, Study 1. *Note. Error bars represent ±one standard error of the mean. For contrasts,* *p < .05, two-tailed.*

### 2.2.2. Generalizations to frogs

Although younger children did not generalize more from diverse than non-diverse samples, their responses did show evidence of systematic category-based induction. For each age group, we conducted a 2 (sample: non-diverse, diverse) × 2 (target: birds, frogs) repeated measures analysis of variance. For each age, children generalized more to birds than frogs (ages, 5–6, $F(1,25) = 4.31$, $p = .05$; ages, 7–8, $F(1,22) = 10.51$, $p = .004$, ages 9–10, $F(1,21) = 19.11$, $p < .001$); see Fig. 4.

Considering responses to the questions about frogs also allows us to determine if children aged 9–10, who engaged in diversity-based reasoning, did so systematically (see Lawson & Fisher, 2011). Indeed, children aged 9–10 generalized from diverse samples more to birds ($M = 2.74$, $SE = .17$) than to frogs ($M = 1.68$, $SE = .27$), $t(21) = 4.05$, $p = .001$. Also, unlike older children's generalizations to birds, their generalizations to frogs did not depend on the diversity of the sample, (Non-diverse = 1.77, $SE = .34$; Diverse = 1.68, $SE = .27$), ns. Thus older children used diversity to guide generalizations to category members, but not to non-category members.

### 2.2.3. Diversity ratings

Younger children did not fail to consider sample diversity because they failed to perceive it. Children's responses to the "do these birds look pretty similar or pretty different?"[1] questions were analyzed via binomial regression models predicting the probabilities of responding that the birds in the sample appeared "pretty different," with sample as a within-subjects factor. Children of each age viewed the diverse samples as more diverse than the non-diverse samples (see Table 2; ages 5–6, Wald $\chi^2(1) = 73.16$, $p < .001$; ages 7–8, Wald $\chi^2(1) = 35.98$, $p < .001$; ages 9–10, Wald $\chi^2(1) = 32.51$, $p < .001$). Inspection of the means (see Table 2) also suggested that responses to the diverse sample varied by age. Indeed, a series of pairwise comparisons revealed that children aged 5–6 rated the diverse samples as significantly more diverse than did children aged 7–8 or aged 9–10, $ps < .001$. Responses to the non-diverse samples did not vary by age. Note that this is exactly opposite to the pattern that would be found if perceptions of diversity explained why younger children did not incorporate diversity into their sample evaluations—younger children rated the diverse samples as highly diverse, yet did not draw broader generalizations from these samples.

### 2.2.4. Control questions

The control questions provide further evidence that children of each age could respond appropriately to the response scale. Children of each age gave significantly higher responses to the control

---

[1] To confirm that asking children to rate the similarity/diversity of the stimuli did not interfere with younger children's performance, an additional 27 younger children (14 male, 13 female, *M* age = 5.14) participated in a follow-up control study. Procedures were identical, with the exception that this question ("Do they look pretty similar to each other or pretty different?") was omitted. Effects were identical to Study 1; there were no effects of sample on these younger children's generalizations (diverse, *M* = 2.82, *SE* = .19; Non-diverse, *M* = 2.96, *SE* = .17), ns.
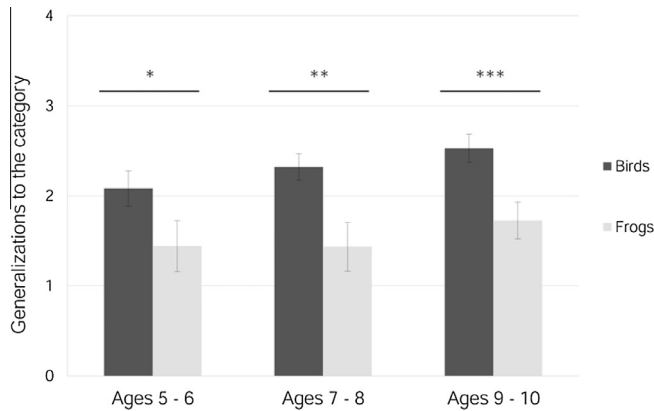
**Fig. 4.** Children's generalizations to birds and frogs, by age, collapsed across samples, Study 1. *Note. Error bars represent ±one standard error of the mean. For contrasts, *p < .05, **p < .01, ***p < .001, two-tailed.*

question meant to elicit responses at the higher end of the scale than those meant to elicit responses towards the lower of the scale (ages 5–6, $t(25)$ = 8.50, $p < .001$; ages 7–8, $t(22)$ = 15.20, $p < .001$, ages 9–10, $t(21)$ = 11.27, $p < .001$), see Table 3.

### 2.2.5. Evidence selection

Children's selections of diverse samples on the evidence selection task showed a similar pattern. Selections of diverse samples were correlated with participant age, $r = .32$, $p = .006$. Binomial regression models revealed that children aged 5–6 selected diverse and non-diverse samples equally often ($M = .50$, $CI = .41$–$.59$), as did children aged 7–8 ($M = .52$, $CI = .42$–$.62$). Children aged 9–10, however, selected diverse samples more often than expected by chance ($M = .70$, $CI = .60$–$.79$; Wald $\chi^2$ (1) = 13.84, $p < .001$).

### 2.3. Discussion

Study 1 used a new paradigm to reveal the same slow pattern of the development of diversity-based reasoning that has been found in much previous work—children aged 5–8 generalized equivalently from non-diverse and diverse samples, whereas children aged 9–10 generalized more from diverse samples. Children of this older age group were also the only ones who reliably sought out diverse samples of evidence to test hypotheses about animal categories.

Despite younger children's failure to consider sample diversity, the task was not overwhelmingly challenging for them. Children in the youngest age group reliably rated the diverse samples as more diverse than the non-diverse samples, indicating that they could readily *perceive* the level of diversity within each sample. Children also responded appropriately using the scale (as shown by the control questions), and generalized more to birds than to frogs. Yet, despite their abilities to manage the task, they did not incorporate sample diversity into their inferences. Further evidence that this task is sufficiently sensitive to assess the abilities of younger children is presented in Studies 2-4.

## 3. Study 2

Study 2 examined children's category-based inferences involving a novel category ("modies"). If younger children's failure to consider sample diversity in Study 1 reflects fundamental developmental differences in the mechanisms available to support induction, then we should find the same slow pattern of development in Study 2. In contrast, if children's prior category knowledge about birds interfered with their diversity-based reasoning in Study 1, then we might find earlier success in Study 2.

**Table 2**
Proportion of samples evaluated as diverse by sample type and age group. CIs in parentheses.

|  | 5–6 year olds | 7–8 year olds | 9–10 year olds |
|---|---|---|---|
| *Study 1* |  |  |  |
| Diverse samples | 0.94 (0.87–0.97) | 0.77 (0.60–0.88) | 0.64 (0.48–0.77) |
| Non-diverse samples | 0.11 (0.05–0.24) | 0.04 (0.01–0.11) | 0.04 (0.01–0.12) |
| *Study 2 (animal condition)* |  |  |  |
| Diverse samples | 0.83 (0.61–0.94) | 0.75 (0.58–0.87) | 0.78 (0.57–0.90) |
| Non-diverse samples | 0.11 (0.04–0.29) | 0.07 (0.02–0.19) | 0.02 (0.00–0.14) |
| *Study 2 (artifact condition)* |  |  |  |
| Diverse samples | 0.75 (0.58–0.87) | 0.73 (0.56–0.85) | 0.73 (0.54–0.86) |
| Non-diverse samples | 0.05 (0.01–0.19) | 0.03 (0.00–0.18) | 0.09 (0.03–0.24) |
| *Study 3 (bird condition)* |  |  |  |
| Diverse samples | 0.84 (0.69–0.93) |  |  |
| Non-diverse samples | 0.28 (0.17–0.42) |  |  |
| *Study 3 (modie condition)* |  |  |  |
| Diverse samples | 0.75 (0.63–0.84) |  |  |
| Non-diverse samples | 0.14 (0.07–0.26) |  |  |
| *Study 4 (bird condition)* |  |  |  |
| Diverse samples | 0.89 (0.77–0.96) |  |  |
| Non-diverse samples | 0.17 (0.09–0.28) |  |  |
| *Study 4 (modie condition)* |  |  |  |
| Diverse samples | 0.79 (0.64–0.89) |  |  |
| Non-diverse samples | 0.11 (0.04–0.25) |  |  |

**Table 3**
Mean responses to the control questions, by age group. SDs in parentheses.

|  | 5–6 year olds | 7–8 year olds | 9–10 year olds |
|---|---|---|---|
| *Study 1* |  |  |  |
| Control question 1 | 0.23 (0.51) | 0.13 (0.34) | 0.27 (0.63) |
| Control question 2 | 2.23 (1.07) | 1.96 (0.56) | 2.41 (0.59) |
| *Study 2* |  |  |  |
| Control question 1 | 0.58 (1.17) | 0.26 (0.67) | 0.30 (0.75) |
| Control question 2 | 2.42 (1.44) | 3.10 (1.30) | 3.20 (1.30) |

Control question 1 is meant to elicit responses toward the low end of the scale. Control question 2 is meant to elicit responses toward the midpoint of the scale in study 1 and the high end of the scale in study 2.

### 3.1. Methods

#### 3.1.1. Participants

Study 2 included 105 children (70 female, 35 male).[2] Five additional children began testing but were excluded from analyses: three because they did not complete the study, one for parental interference, and one for experimenter error. Given the findings of Study 1 and of prior work, we considered the children as three age groups: 33 children aged 5–6 (*M* age = 6.2), 42 children aged 7–8 (*M* age = 7.9), and 30 children aged 9–10 (*M* age = 9.7). Because some researchers have suggested that diversity-based reasoning might be easier for young children in the case of artifact (instead of animal) categories (Heit & Hahn, 2001; Shipley & Shepperson, 2006), participants were randomly assigned to hear modies described as

---

[2] All children in the target age range visiting the museum during the times in which we collected data were invited to participate. In Study 2 only, this recruitment strategy resulted in a gender imbalance. Male and female participants showed very similar patterns. Both groups showed an effect of sample, (Males, $F(1,29) = 9.20$, $p = .005$; Non-diverse, $M = 1.76$, $SE = .13$, Diverse, $M = 2.42$, $SE = .16$; Females, $F(1,64) = 8.63$, $p = .005$, Non-diverse, $M = 1.83$, $SE = .12$, Diverse, $M = 2.18$, $SE = .13$), and no main or interactive effects of age or category-type. Thus, this gender imbalance apparently did not influence our findings.
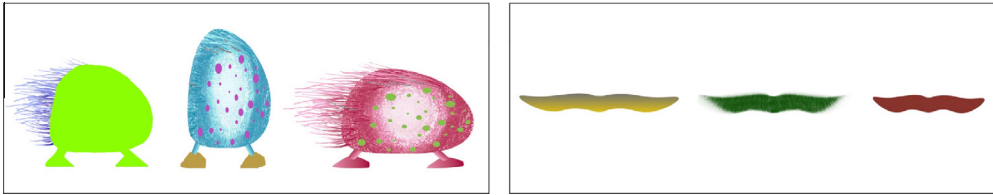
**Fig. 5.** Stimuli used to introduce the target categories, Study 2 and "modie" condition of Study 4 (left) and Study 3 (right).

animals or artifacts; age groups were set in advance so that equal numbers of children at each age could be randomly assigned to each domain condition.

### 3.1.2. Scale Training and Category Introduction

First, children completed the same training on the dot scale as in Study 1. Then they were introduced to the novel category, modies. To do so, children were shown three modies (see Fig. 5), which were perceptually identical across conditions but were described as either artifacts or animals (see Brandone & Gelman, 2009; Rhodes, Gelman, & Karuza, 2014, for evidence that children will interpret these stimuli as either animals or artifacts depending on the accompanying description) as follows:

"I'm going to show you some modies. Here is a modie."
[*Artifact*] "Modies are invented by people in a place far, far away. People who live far away build modies and they use modies a lot."
[*Animal*] "Modies are born in a place far, far away. Modies grow up and live in this place far away."
"Here is another modie. And here is another modie. Remind me, what are these called?"
[*Artifact*] "Modies are invented, built, and used by people in a place far, far away."
[*Animal*] "Modies are born, grow up, and live in a place far, far away."
"I'm going to ask you to make some guesses about modies."

### 3.1.3. Induction task

Next, a collage showing many modies was brought out (see Fig. 2) and children were told, "See all these modies? We have a lot of different modies here. I'm going to tell you some special things about some of these modies, and your job is to guess how many of the other modies have those things too." This collage remained on the table throughout testing (as had the bird collage in Study 1).

Subsequently, children completed induction trials, following an identical structure to those in Study 1 (see Fig. 1). The list of properties asked about is given in Table 1. All procedures for blocking and counterbalancing were identical to Study 1, and children completed two control questions at the end (designed to elicit responses at either the lower or higher end of the scale).

### 3.1.4. Manipulation check

To confirm that children understood the domain manipulation, at the end of the experiment, children were asked, "Is a modie the kind of thing that eats food, or the kind of thing that runs on batteries?" (scored 1 = food, 0 = batteries).

### 3.1.5. Participant background questionnaire

Because children for this research were recruited at a major cultural institution (the American Museum of Natural History in New York City), we anticipated some variability in our sample with respect to children's experiences with nature. As such experiences have been found to relate to children's biological concepts in prior work (Coley, Hayes, Lawson, & Moloney, 2004; Medin, Waxman, Woodring, & Washinawatok, 2010; Ross, Medin, Coley, & Atran, 2003), parents of participating children were asked to fill out a questionnaire asking for additional information about their family.

Of the 105 participants, 59 families chose to provide this additional information. The questionnaire asked parents to identify their home state and describe their community as urban, suburban, or rural; to indicate whether their family has pets; to rate (on a 6 point scale) how often their child engaged in a series of activities in the past year, including camping, hiking, fishing, gardening, taking care of pets, taking care of farm animals, and visiting zoos; to rate (on a 4 point scale) how much education about birds their child had received, their child's interest in nature, and their child's interest in reading books about nature and animals; and how many times in the previous year their child had visited the American Museum of Natural History.

## 3.2. Results

Of the families who provided additional background information, 36 lived in urban environments, 21 in suburban, and 2 in rural; 39 were from New York, 10 were from New Jersey, and 9 were from other states. We conducted a series of preliminary analyses testing for effects of the child's home community, whether the child had pets, each item on the 6-point "activities in the past year" scale (as well as a composite), each item on the 4-point "interest in nature" scale (as well as a composite), and number of visits to the museum, on children's generalizations from diverse and non-diverse samples (also including age-group and category-type in each analysis). None of the background variables significantly predicted performance or interacted with any of our key independent variables. Therefore, we did not consider these participant variables further.

### 3.2.1. Generalizations to "Modies"
Repeated measures analyses of variances tested for main and interactive effects of sample (diverse vs. non-diverse), category-type (animal vs. artifact), and age group (ages 5–6, 7–8, 9–10). This analysis revealed only a main effect of sample, $F(1,99) = 17.60$, $p < .001$. Children generalized more from diverse ($M = 2.25$, $SE = .10$) than non-diverse ($M = 1.80$, $SE = .09$) samples, Cohen's $D = .44$. Although there were no effects of age in the main analysis, to confirm that the same pattern was found across age (particularly since this finding differed from Study 1 and from prior work), we conducted follow-up analyses separately within each age group. Indeed, children's responses followed the same pattern at each age group (ages 5–6, Diverse = 2.26, $SD = 1.03$, Non-diverse = 1.87, $SD = .93$, $t(32) = 2.34$, $p = .02$, Cohen's $D = .40$; ages 7–8, Diverse = 2.11, $SD = .99$, Non-diverse = 1.81, $SD = .89$; $t(41) = 1.67$, $p = .1$, Cohen's $D = .32$; ages 9–10, Diverse = 2.28, $SD = 1.01$, Non-diverse = 1.66, $SD = .94$, $t(29) = 3.31$, $p = .003$; Cohen's $D = .74$; all tests two-tailed), though this contrast did not reach significance among children ages 7–8. Further follow-up analyses confirmed no interactions between category-type and sample in any age group examined separately.

### 3.2.2. Diversity ratings and control questions
Children were more likely to respond that the diverse samples ($M = .76$, $CI = .69–.82$) than the non-diverse samples ($M = .05$, $CI = .03–.10$) appeared "pretty different from each other," Wald $\chi^2 (1) = 140.12$, $p < .001$. There were no main or interactive effects of age or category-type on these judgments, see Table 2. Children of each age responded appropriately to the control questions (see Table 3, ages 5–6, $t(32) = 5.88$, $p < .001$; ages 7–8, $t(41) = 13.16$, $p < .001$; ages 9–10, $t(29) = 11.16$, $p < .001$).

### 3.2.3. Manipulation check
Despite the lack of domain effects on children's responses, children did indeed understand the domain manipulation as intended. Children were more likely to say that modies in the animal than artifact condition "ate food" (instead of ran on batteries); $M$ animal = .63, $CI = .48–.76$, $M$ artifact = .38, $CI = .26–.51$ Wald $\chi^2 (1) = 6.09$, $p = .01$. Further, we reanalyzed the central data via a 3 (age-group) × 2 (sample) × 2 (domain) repeated measures analysis of variance with responses to this question marking domain instead of randomly assigned condition. This analysis revealed the main effect of sample found in the central analyses, but again, found no main or interactive effects of domain or age.

### 3.3. Discussion

In contrast to Study 1, Study 2 found no evidence of developmental change—across age, children generalized more broadly from diverse than non-diverse samples of evidence. This is the first study to provide clear evidence of diversity-based reasoning in young children (ages 5–6). The tasks used in Studies 1 and 2 were structurally identical, but varied in whether they tested how children reasoned about a familiar biological kind or a novel category for which they held no prior knowledge. Thus, these findings run counter to the explanation that children's failure in Study 1 (and in prior work) reflects developmental differences in the mechanisms available to support induction, and instead suggest developmental differences in when children access those mechanisms.

## 4. Study 3

The difference in children's performance across Studies 1 and 2 could suggest that whether young children engage in diversity-based reasoning depends on the availability of prior category knowledge, as we proposed. These differences across studies, however, could also reflect several methodological differences between Studies 1 and 2. First, different perceptual stimuli were used across the two studies. Thus, it is possible that the perceptual stimuli used in Study 2 somehow facilitated diversity-based reasoning. For example, the more abstract nature of the stimuli presented in Study 2 may have led children to focus more on perceptual variation. Second, different properties were asked about in the test questions used in Studies 1 and 2 (see Table 1), with those in Study 1 reflecting more internal, biological qualities. Changing these properties was necessary for Study 2 so that the same properties could be asked about for animals and artifacts, but by consequence, the properties in Study 1 were more biological, whereas those in Study 2 were more behavioral. Thus, one possibility is that children fail to engage in diversity-based reasoning for internal, biological properties, but do so for more behavioral qualities. To address these concerns, Study 3 compared children's reasoning about a familiar natural kind (birds) and a novel category (modies) using identical perceptual stimuli and properties in the test questions. Because our primary interest here was on the variable performance of the youngest children across the two previous studies, Study 3 focused on children ages 5–6.

### 4.1. Methods

Participants included 58 children ages 5.0–6.9 (*M* age = 6.0, 30 male, 28 female). Five additional children began testing but were excluded (2 because they did not complete the test questions, 1 for an experimenter error, and 2 for failing to understand the scale during training). Of the 58 participating children, 53 parents filled out the supplementary background questionnaire used in Study 2. These data showed that 33 participants lived in urban environments, 19 in suburban, and 1 in rural; 35 participants were from New York, 6 from New Jersey, 2 were from Canada, and 9 were from other states or countries. We examined all data from this questionnaire as in Study 2, and again found that none of the background variables predicted performance or interacted with condition. These data were not considered further.

#### 4.1.1. Scale training and category introduction

Participants were randomly assigned to the bird or modie condition. Procedures were very similar to Studies 1 and 2 regarding training with the scale, introduction of the task, and the structure of the test questions (see Fig. 1). The critical exception was in the introduction to the category. In both conditions, children were shown identical perceptual stimuli (see Figs. 2 and 5).

In the modie condition, these stimuli were introduced as follows: "I'm going to show you some modies. Here is a modie. Modies are born in a place far, far away. Modies grow up and live in this place far, far away. Here is another modie. And here is another modie. Remind me, what are these called? Yes – modies are born, grow up, and live in a place far, far away." In the bird condition, the same stimuli were introduced: "I'm going to show you some birds. Here is a bird. Birds come out of eggs. Birds grow up and live on our planet. Here is another bird. And here is another bird. Remind me, what are these

called? Yes – birds hatch, grow up, and live on our planet." Thus, what differed across conditions was only whether the children were asked to consider the stimuli as members of a familiar natural kind (birds) or as a fictional animal category (modies).

### 4.1.2. Induction task, control questions, and manipulation check

The test questions were identical across conditions, followed the same structure as the previous studies, and asked about the same internal, biological properties as in Study 1 (with one exception, see Table 1). After the main test questions, children completed control questions (one that presented a diverse sample and one that presented a non-diverse sample) that described incidental properties, for which generalization would be inappropriate (e.g., "These three birds/modies fell down and got a little scrape last week." "These three birds/modies ate something funny and got a tummy ache a couple of days ago"). Also, to confirm that children understood the conditions as intended, at the end of the experiment, children in the modie condition were asked, "What do you think a modie is?" Children in the bird condition were asked, "What do you think these are?"

### 4.2. Results and discussion

We ran a 2 (sample) × 2 (category-type) repeated measures analysis of variance on children's generalizations following diverse and non-diverse samples. These analyses revealed a sample × category-type interaction, $F(1, 56) = 5.14$, $p = .03$. As shown in Fig. 6, children generalized more from diverse samples than non-diverse samples only when stimuli were referred to as "modies," $p = .02$, Cohen's $D = .43$. To test whether this effect varied across age within this sample of younger children, we re-ran this analysis including children's exact age as a continuous predictor; this analysis replicated the sample × condition interaction, $F(1, 54) = 4.70$, $p = .04$, and found no main or interactive effects of age ($ps > .1$). Thus, when presented with identical perceptual stimuli, children aged 5–6 engaged in diversity-based reasoning when they thought of the stimuli as members of a novel animal category, but not when those same stimuli were presented as members of a familiar kind.

As in previous studies, children were more likely to respond that the diverse samples ($M = .80$, $CI = .71–.87$) than the non-diverse samples ($M = .20$, $CI = .13–.29$) appeared "pretty different from each other," Wald $\chi^2 (1) = 73.64$, $p < .001$. There were no main or interactive effects of category-type on these judgments (see Table 2). Further, analysis of the control questions revealed no main or interactive effects of condition or sample ($ps > .5$). Overall children tended not to generalize the incidental properties to many other category members ($M = 1.17$, $SE = .13$). Confirming that children understood the conditions as intended, when asked to identify the stimuli, all of the children in the bird condition responded with "birds," whereas in the modie condition, none of the children identified the stimuli as birds. Responses varied; for example, children responded with: "a kind of magical creature," "an animal," "like a worm."

The data from Study 3 confirm that children ages 5–6 engage in diversity-based reasoning for novel, but not familiar, animal categories—even when the same perceptual stimuli are used to depict both types of categories and when children are asked in both conditions to make inferences regarding internal, biological properties. Thus, these data conceptually replicate the differences found across Studies 1 and 2 within a single sample of children, and suggest that these differences reflect meaningful differences in how children reason about familiar and novel categories.

## 5. Study 4

Study 4 directly tests the possibility that younger children's category knowledge—in particular, their knowledge of the typical examples and properties—interferes with their diversity-based reasoning. If so, when children have such knowledge, as in the case of familiar natural kinds, they would treat diverse and non-diverse samples as equivalently informative for category-wide generalizations when typicality is held constant across the samples. They should, however, draw broader generalizations from typical than atypical samples for these categories (see López et al., 1992; Rhodes, Brickman, et al., 2008). In contrast, when children lack this knowledge, as in the case of novel categories, they
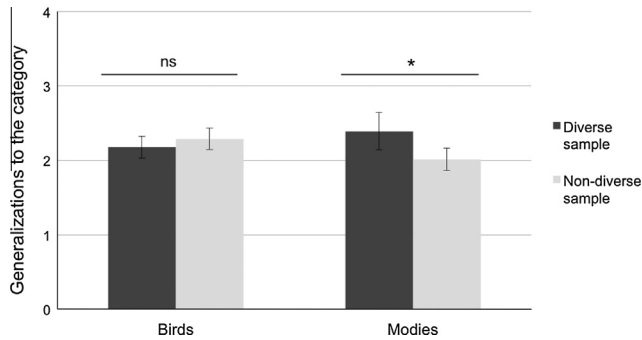
**Fig. 6.** Children's generalizations from diverse and non-diverse samples, by category-type, Study 3. _Note. Error bars represent ±one standard error of the mean. For contrasts, ∗p < .05, two-tailed._

could adopt a coverage-based strategy instead, which includes a preference for diverse—but not necessarily for typical—samples.

For this study, children were randomly assigned to "bird" or "modie" conditions. We assessed the structure of their typicality beliefs and then assessed their diversity-based and typicality-based reasoning. We expected to find that children generalize more from typical than atypical samples for birds (but not modies), but from diverse more than non-diverse samples for modies (but not for birds).

### 5.1. Methods

Participants included 44 children ages 5.0–6.9 (_M_ age = 6.1, 14 male, 30 female). Three additional children began testing but were excluded from analyses; 1 for not completing the test questions and 2 for familial interference.

#### 5.1.1. Procedures for typicality task
First, children were trained on the task for assessing typicality. These methods were based on Rhodes and Gelman (2009), and began with an unrelated example about sports:

"Let's pretend we're talking to someone who doesn't know anything about sports. He doesn't even know what a sport is at all! We need to show him some activities so that he'll really understand what sports are like. Let's pick some activities that are really good ones to show what a sport really is. You can tell me: "Thumbs up: Yes! That is a really good one to show what a sport is really like. Thumbs to the side: Sort of! That is sort of a good one show what a sport is really like. Or Thumbs down: No! That is not a good one to show what a sport is like at all."

Children were presented with a card that showed a cartoon thumb in each of the three positions; the experimenter pointed to the corresponding picture and also demonstrated with his or her own thumb to introduce the scale. Children could respond to questions verbally, by pointing to a picture on the scale, or by using their own thumb.

Children then completed three practice trials:

"This is baseball. Is baseball a good one to show what a sport is really like? (Children were expected to respond "thumbs up" or "yes"). This is reading. Is reading a good one to show what a sport is really like? (Children were expected to respond "thumbs down" or "no"). This is hopscotch. Is hopscotch a good one show what a sport is really like? (Children were expected to respond "thumb to the side" or "sort of").

These practice trials were intended to provide experience using each of the three scale points; thus if the child gave an unexpected answer, the experimenter responded, for example: "That's a good idea, but I think hopscotch is _sort of_ a good one to show what a sport is really like. Can you point to the thumb that shows sort of?"

The main typicality trials assessed children's beliefs about the typicality of exemplars of birds or modies. Participants were randomly assigned to bird or modie conditions. The stimuli were the same as in Study 1 (for the bird condition) and Study 2 (for the modie condition), and children were introduced to the categories as in Study 3. Thus, they were presented with the sheets containing many category exemplars shown in Fig. 2, and heard the scripts used in Study 3 to describe the categories. They were then asked to evaluate the typicality of individual exemplars of birds or modies, for example: "Now let's pretend we're talking to someone who doesn't know anything about birds. He doesn't even know what a bird is at all! We need to show him some birds so that he'll know what birds are really like. I'm going to show you some things. For each one, you can use the thumbs to tell me, "Yes, that is a good one to show what a bird is really like," "Sort of; that is sort of a good one to show what a bird is really like," or "No, that is not a good one to show what a bird is like at all."

Children completed 18 typicality trials. Each trial presented a single exemplar. The trials included 6 typical category members (for birds: three bluebird trials and three robins trials; for modies: 6 exemplars that shared all of the most common properties of modies shown in the collage in Fig. 2), 6 atypical category members (for birds: three peacock trials and three penguin trials; for modies, 6 exemplars with two unusual features—for example, these modies did not have hair and were of a color not included in the collage in Fig. 2), and 6 non-members (for birds: three butterfly trials and three flying squirrel trials; for modies, 6 novel exemplars that did not share color, shape, or any other systematic features with modies in the collage in Fig. 2; for illustrations, see the online supplementary materials). For birds, these exemplars were chosen based on Rhodes & Gelman, 2009 and Rhodes, Brickman, et al. (2008), which found that children of this age (as well as adults) assigned these typicality levels to these exemplars. For each item, children were shown a single picture and asked, "Is this one a good one to show what a [bird/modie] is really like?" Children responded verbally or by using the thumb scale ("yes" = 2, "sort of" = 1, or "no" = 0).

Note that for modies, it was possible (and straightforward) to draw systematic perceptual typicality judgments simply by comparing each exemplar shown during the task to the collage of modies shown in Fig. 2, which remained on the table throughout testing. Thus, our prediction was not that children would show systematic typicality beliefs for birds but not for modies. Rather, we included the typicality task mainly to confirm that children did indeed assign the anticipated levels of typicality to these exemplars. Despite being able to recognize the perceptual prototype of modies, however, this prototype would not hold the same conceptual weight for modies as for birds. Since children have limited experience with modies, they cannot have clear ideas about how these perceptual features correspond to the deeper biological or behavioral properties that are asked about in the test questions. In other words, for modies, children lack the prior conceptual knowledge of category-typical biological or behavioral features that would be necessary to view typical exemplars as informative for drawing deeper inferences.

### 5.1.2. Procedures for scale training and induction task

After the typicality task, procedures were identical to Study 1 (Bird condition) and Study 2 (Modie condition) for scale training, introduction of the induction task, and the structure of the induction test questions (see Fig. 1). All test questions featured the more biological properties asked about in Studies 1 and 3 (have hollow bones, don't have any teeth, have a syrinx inside, see special colors that people can't see, have air sacs inside, make a special chemical called calcium). Children completed three test questions for non-diverse samples and three for diverse samples, in counter-balanced order across participants. Subsequently, to assess typicality-based generalizations, children completed six induction trials for single exemplars: three that presented typical exemplars and three that presented atypical exemplars, in counter-balanced order across participants. These items presented new exemplars drawn from the categories used in the typicality assessment task (e.g., in the bird condition, new exemplars of robins, bluebirds, peacocks, and penguins; in the modie condition, new exemplars of modies with similar features as the typical and atypical modies shown during the typicality assessment task). These questions also asked about biological properties, but different ones than were used for the diverse and non-diverse induction trials (has a special thing inside called a furcula, has little scales on it, has a four-chambered heart, digests food very quickly, has a bony septum, makes a chemical called uric acid); properties rotated across typical and atypical samples across participants.

A sample question is: "Here is a bird (*show a picture of a penguin*). This bird has a special thing inside called a furcula. This bird has a furcula inside. Now tell me your best guess, of all the birds in the world, how many birds have a furcula?"

In the modie condition, at the end of the experiment children were asked, as in Study 2, "Is a modie the kind of thing that eats food or runs on batteries?" to confirm that they viewed modies as living kinds. As intended, this question confirmed that children reliably construed modies as a living kind (73% of children responded that modies "eat food" rather than "run on batteries", $p = .05$, binomial test).

## 5.2. Results and discussion

Analyses proceeded in four phases. First, we tested whether typicality ratings varied as intended by exemplar-type. Second, we tested whether we replicated the findings from Studies 1–3 regarding diversity-based reasoning—in which case children in the modie condition, but not the bird condition, should generalize more from diverse than non-diverse samples. Third, we tested the prediction that children in the bird condition, but not the modie condition, would generalize further from typical than atypical exemplars. Finally, we directly examined how children's typicality ratings relate to their diversity-based reasoning.

### 5.2.1. Typicality ratings

A 2 (condition: bird, modie) × 3 (exemplar-type: typical, atypical, nonmember) repeated measures analysis of variance confirmed that children construed the exemplars as intended in both conditions: Children rated the typical exemplars ($M = 1.68$ (out of 2), $SE = .07$) as more typical than the atypical exemplars ($M = 1.25$, $SE = .09$), which were rated as more typical than the non-members ($M = .27$, $SE = .06$); main effect of type, $F(2,84) = 103.91$, $p < .001$; all contrasts significant at $p < .001$. There were no main or interactive effects of condition ($ps > .50$), see Table 4.

### 5.2.2. Generalizations from diverse and non-diverse samples

Replicating our previous studies, a 2 (condition: bird, modie) × 2 (sample: diverse, non-diverse) repeated measures analysis of variance revealed a sample × condition interaction (see Fig. 7), $F(1,42) = 4.10$, $p < .05$. In the bird condition, children's generalizations did not differ by sample diversity ($p > .6$), but in the modie condition, children generalized further from diverse than non-diverse samples, $p = .002$, Cohen's $D = .55$. As in Study 3, including participants' exact age as a continuous predictor revealed no main or interactive effects of age, and the sample × condition interaction replicated in this analysis, $F(1,41) = 3.98$, $p = .05$.

As in previous studies, children were more likely to respond that the diverse samples ($M = .85$, $CI = .76–.91$) than the non-diverse samples ($M = .13$, $CI = .08–.22$) appeared "pretty different from each other," Wald $\chi^2 (1) = 85.16$, $p < .001$. There were no main or interactive effects of category-type on these judgments (see Table 2).
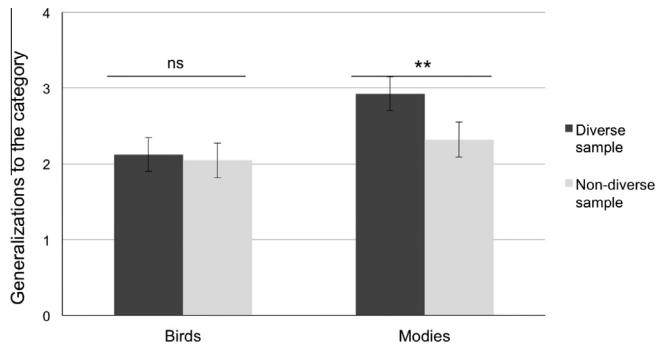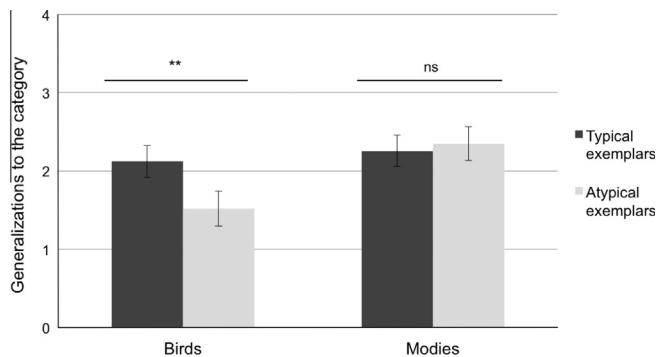
### 5.2.3. Generalizations from typical and atypical exemplars

As predicted, a 2 (condition: bird, modie) × 2 (typicality: typical, atypical) repeated measures analysis of variance revealed a typicality × condition interaction, $F(1,39) = 5.41$, $p = .03$, see Fig. 8. In the bird condition, children generalized further from typical than atypical exemplars, $p = .008$, Cohen's $D = .81$, whereas in the modie condition, generalizations did not vary by sample typicality, $p > .6$. Thus, although children in the modie condition drew typicality distinctions based on perceptual features (as shown in the analysis of their typicality structure above), they did not view typical exemplars as more informative for drawing inferences about the kind in the case of a novel category, suggesting that what matters for guiding these inferences is children's knowledge of the deeper features associated with the category (e.g., flying, digs for food, and so on), not their ability to recognize perceptually typical members.

**Table 4**
Mean (SE) typicality ratings by condition and exemplar type.

|  | Typical | Atypical | Non-member |
|---|---|---|---|
| Bird | 1.74 (.10) | 1.23 (.12) | .29 (.09) |
| Modie | 1.61 (.10) | 1.27 (.12) | .25 (.09) |



**Fig. 7.** Children's generalizations from diverse and non-diverse samples, by category-type, Study 4. *Note. Error bars represent ±one standard error of the mean. For contrasts,* **p* < .01, two-tailed.



**Fig. 8.** Children's generalizations from typical and atypical exemplars, by category-type, Study 4. *Note. Error bars represent ±one standard error of the mean. For contrasts,* **p* < .01, two-tailed.

### 5.2.4. Relations of typicality ratings to generalizations

To test directly how children's typicality beliefs related to their generalizations from diverse and non-diverse samples, we reran the 2 (condition: birds, modies) × 2 (sample: diverse, non-diverse) repeated measures analysis of variance, including children's typicality ratings for the typical exemplars as a continuous predictor. This analysis indeed yielded a sample × typicality rating interaction, $F(1,41) = 4.77$, $p = .04$. To determine the nature of this effect, we examined the correlation between children's typicality ratings of the typical exemplars and a difference score that reflected the extent to which they drew broader inferences from diverse than non-diverse samples (generalizations from diverse–generalizations from non-diverse); this correlation was negative, $r = -.34$, $p = .03$. Thus, the more firmly that children viewed the typical exemplars as the best way to illustrate the kind on the typicality task, the less likely they were to engage in diversity-based reasoning.

We reran this analysis using instead children's ratings of the *atypical* exemplars as a predictor and found no effects of these beliefs. Thus, it was not the case that children were more likely to engage in

diversity-based reasoning to the extent that they view atypical exemplars as informative, *rather* the extent to which they focused on the highly typical members made them less likely to do so.

Although children recognized the typical exemplars as highly typical for both birds and modies (see Table 4), there was some evidence that they held clearer prototypes for the familiar category. In particular, the average inter-item correlation among children's ratings of each typical exemplar was significantly higher in the bird condition ($r = .68$, range = .42–.85) than in the modie condition ($r = .48$, range = $-.08$ to .79), $p = .01$. Thus, children held more consistent beliefs regarding which exemplars best represented the kind for familiar categories.

## 6. General discussion

Determining how much to generalize information obtained from limited samples is a critical part of efficient knowledge development. Here we have shown that the processes available to support this component of induction are continuous across childhood (from ages 5 to 10), but that when children access these mechanisms varies across age. For familiar natural kind categories—the focus of much previous research on category-based induction (Gelman, 2003; Heit, 2000; Medin et al., 2003; Osherson et al., 1990)—young children treat non-diverse and diverse samples as equivalently informative for drawing inferences about broader kinds. Only at ages 9–10 do children generalize more from diverse samples (as is found among adults, Osherson et al., 1990 and predicted by normative models, Heit, 2000). Yet, for novel animals and artifacts, younger children successfully engaged in diversity-based reasoning. Study 3, which used identical perceptual stimuli to depict a familiar animal category and a novel one, confirms that the youngest children's differential responding to familiar and unfamiliar kinds stems from the conceptual knowledge they bring to the task. Study 4 replicated this effect and indicated that the centrality of typical examples in young children's representations of familiar biological kinds interferes with their diversity-based reasoning for these categories.

Previous work found that although young children have the basic ability to track diversity (Heit & Hahn, 2001; Rhodes, Gelman, et al., 2008; Rhodes et al., 2010; Shipley & Shepperson, 2006; Zhong et al., 2014) and reason about the relation of diversity within a sample to diversity within a population (Xu & Garcia, 2008; Xu & Tenenbaum, 2007), they often fail to incorporate such information into their category-based induction (Carey, 1985; Gutheil & Gelman, 1997; López et al., 1992; Rhodes, Gelman, et al., 2008, 2010). The present studies 2–4 are the first to clearly demonstrate that children ages 5–6 generalize information more broadly from diverse than non-diverse samples. As revealed by the control questions in Study 3, when these younger children did consider sample diversity, they did so in a reasonable and adult-like manner—to shape their inferences about biological but not incidental properties. Thus, these studies are the first to reveal normative diversity-based reasoning in children below the age of 8.

As young children engaged in diversity-based reasoning for novel kinds in Studies 2–4, their failure to do so for familiar biological kinds in these studies and in prior work is even more striking and important to explain. In the present studies, children engaged in diversity-based reasoning for novel but not for familiar kinds even when presented with identical tasks, perceptual stimuli (Study 3), and test properties (Studies 3–4). Thus, comparisons both across and within our studies suggest that the differences in children's reasoning about familiar and novel kinds cannot be explained by differences in task demands or computational requirements.

Instead, we propose that prior category knowledge—particularly knowledge of category-typical examples and features—interferes with younger children's diversity-based reasoning for familiar biological kinds. In Study 4, children demonstrated typicality-based reasoning for these categories—they generalized to more birds from robins than from penguins, for example. Although children recognized the perceptual typicality structure of the novel category (as shown by their typicality ratings), they did not incorporate typicality information into their category-based induction for these categories (relying instead on sample diversity). This pattern suggests that young children have access to both diversity-based and typicality-based reasoning, but rely on them in different contexts (unlike adults, who consider both diversity and typicality information simultaneously, Osherson et al., 1990).

According to Osherson et al. (1990), when adults are asked to consider the extent to which a sample is informative regarding *birds*, for example, they bring to mind their concept of bird, which

includes extensive variability (including typical examples such as robins and bluebirds, as well as less typical examples such as penguins and ostriches), and seek out a sample that covers this broad category. More diverse samples clearly provide better coverage. Typical samples do as well, because an exemplar's average similarity to all other category members is assumed to be a strong determinant of typicality. In contrast, we suggest that when younger children are asked to generalize to *birds*, they bring to mind a less variable set, one composed primarily of typical examples (e.g., robin, bluebird) and not less typical ones (e.g., penguins and ostriches; see Anglin, 1986; Hayes & Taplin, 1993; Mervis & Pani, 1980; Rosner & Hayes, 1977), and then seek out samples that share important features with the typical examples that they have brought to mind (e.g., small birds that fly). Note that this process involves prior conceptual knowledge about birds (e.g., knowing that typical birds fly, dig for food, and so on—properties that cannot be seen in the presented stimuli), not merely the detection of a perceptual prototype. Thus, typical samples are informative to young children not because they provide coverage, but because they share deep features with the prototypical examples they have brought to mind. From this perspective, children do not seek out samples that provide coverage of the broad category of birds because they have not brought highly variable examples to mind (e.g., they have not brought to mind large non-flying birds). Consistent with this interpretation, prior work has shown that when 7-year-olds are first induced to bring to mind variable examples of natural kinds, they seek out diverse samples when they would not otherwise do so (Rhodes & Brickman, 2010). On this account, when children hold less conceptual knowledge about a category, as in the case of novel kinds for which they do not know how perceptual features correspond to deeper properties, they bring to mind the more variable set they have been shown instead of a select few examples, and consequently seek out samples that provide broader coverage.

This explanation leaves open the question of why young children bring to mind less variable examples than older children or adults when asked to reason about familiar natural kinds. One possibility is that young children's representations exclude atypical exemplars (e.g., that they do not recognize penguins or ostriches as birds at all). Alternately, children could be aware of atypical exemplars, but simply not bring them to mind as easily as adults do. Identifying which of these possibilities is the case will be a useful area for future work. Identifying how and when children begin to bring to mind more within-category variability as their concepts develop has important practical implications as well. For example, a tendency to overlook within-category variability interferes with children's understanding of evolution via natural selection (Gelman & Rhodes, 2012; Shtulman & Schulz, 2008). Thus, identifying the extent to which representations of variability or retrieval of variability change across age has key implications for designing developmentally appropriate educational approaches.

These studies were conducted in the context of a large natural history museum in a major metropolitan area. We collected information from participants to see if children's exposure to the museum or their previous experiences with animals or in nature influenced performance, and did not find this to be the case. Nevertheless, these findings do not preclude the possibility that certain populations of younger children—perhaps those with extensive experiences with nature or expertise—might show diversity-based reasoning for natural kinds at a younger age. Examining this possibility by targeting particular populations for future study—including populations that have greater variability in their experiences than those included here—could be another promising way to distinguish the extent to which category-based induction depends on differences in category knowledge (e.g., Coley et al., 2004; Medin et al., 2010).

The present work considered whether the processes by which humans acquire knowledge stay constant across development or undergo fundamental change, using diversity-based reasoning as a test case. Our studies revealed both continuity and change. These studies provide the first clear evidence that children have access to diversity-based reasoning by at least age 5, consistent with the view that the mechanisms that underlie inductive inference show important continuity. Yet, these studies also revealed important developmental change in when children access these mechanisms. Younger children did so for novel categories, but not for familiar natural kinds, suggesting that category knowledge—particularly knowledge of category-typical examples and features—interferes with diversity-based reasoning in early childhood. Examining precisely how this prior knowledge is incorporated into children's and adults' reasoning processes is an important area for future work.

**Table A1**
Mean generalizations from single samples in all studies, by age group. SDs in parentheses.

|                          | 5–6 year olds | 7–8 year olds | 9–10 year olds |
|--------------------------|---------------|---------------|----------------|
| Study 1                  | 1.73 (0.20)   | 1.73 (.21)    | 2.32 (.21)     |
| Study 2                  | 1.89 (0.14)   | 1.67 (0.12)   | 1.51 (0.14)    |
| Study 3 (bird condition) | 1.70 (0.22)   |               |                |
| Study 3 (modie condition)| 2.13 (0.22)   |               |                |

## Acknowledgments

## Appendix A

See Table A1

## Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cogpsych.2015.07.003.

## References

Anglin, J. (1986). Semantic and conceptual knowledge underlying the child's words. In S. A. Kuczaj & M. D. Barret (Eds.), *The development of word meaning* (pp. 83–97). New York: Springer-Verlag.

Atran, S. (1990). *Cognitive foundations of natural history: Towards an anthropology of science*. New York: Cambridge University Press.

Bacon, F. (1898). Novum organum. London: George Bell & Sons (Original work published 1620).

Brandone, A. C., & Gelman, S. A. (2009). Differences in preschoolers' and adults' use of generics about novel animals and artifacts. A window onto a conceptual divide. *Cognition, 110*, 1–22. http://dx.doi.org/10.1016/j.cognition.2008.08.005.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: Bradford Books.

Coley, J., Hayes, B., Lawson, C., & Moloney, M. (2004). Knowledge, expectations, and inductive inferences within conceptual hierarchies. *Cognition, 90*, 217–253. http://dx.doi.org/10.1016/s0010-0277(03)00159-8.

Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: Evidence from 4.5- and 6-month-olds. *Developmental Psychology, 49*, 243–249. <http://psycnet.apa.org/doi/10.1037/a0028278>.

Feeney, A., & Heit, E. (2007). *Inductive reasoning: Experimental, developmental, and computational approaches*. New York: Cambridge University Press.

Feeney, A., & Heit, E. (2011). Properties of the diversity effect in category-based inductive reasoning. *Thinking & Reasoning, 17*, 156–181. http://dx.doi.org/10.1080/13546783.2011.566703.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday life*. New York: Oxford University Press.

Gelman, S. A., & Rhodes, M. (2012). "Two-thousand years of stasis": How psychological essentialism impedes evolutionary understanding. In K. R. Rosengren, S. Brem, E. M. Evans, & G. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning about evolution* (pp. 3–21). New York: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199730421.003.0001.

Gutheil, G., & Gelman, S. A. (1997). Children's use of sample size and diversity information within basic- level categories. *Journal of Experimental Child Psychology, 64*, 159–174. http://dx.doi.org/10.1006/jecp.1996.2344.

Hayes, B. (2007). The development of inductive reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Experimental, developmental and computational approaches* (pp. 25–54). Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511619304.003.

Hayes, B., & Taplin, J. (1993). Developmental differences in the use of prototype and exemplar-specific information. *Journal of Experimental Child Psychology, 55*, 329–352. http://dx.doi.org/10.1006/jecp.1993.1019.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7*, 569–592. http://dx.doi.org/10.3758/bf03212996.

Heit, E., & Hahn, U. (2001). Diversity-based reasoning in children. *Cognitive Psychology, 43*, 243–273. http://dx.doi.org/10.1006/cogp.2001.0757.

Heit, E., Hahn, U., & Feeney, A. (2005). Defending diversity. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the lab: Festschrift in honor of Douglas L. Medin* (pp. 87–99). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/11156-006.

Kemp, C., & Tenenbaum, J. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116*, 20–58. http://dx.doi.org/10.1037/a0014282.

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition, 31*, 155–165. http://dx.doi.org/10.3758/bf03196090.

Lawson, C. (2014). Three-year-olds obey the sample size principle of induction: The influence of evidence presentation and sample size disparity on young children's generalizations. *Journal of Experimental Child Psychology, 123*, 147–154. http://dx.doi.org/10.1016/j.jecp.2013.12.004.

Lawson, C., & Fisher, A. (2011). It's in the sample: The effects of sample size and sample diversity on the breadth of inductive generalization. *Journal of Experimental Child Psychology, 110*, 499–519. http://dx.doi.org/10.1016/j.jecp.2011.07.001.

Li, F., Cao, B., Li, Y., Li, H., & Deák, G. (2009). The law of large numbers in children's diversity-based reasoning. *Thinking & Reasoning, 15*, 388–404. http://dx.doi.org/10.1080/13546780903343227.

López, A. (1995). The diversity principle in the testing of arguments. *Memory & Cognition, 23*, 374–382. http://dx.doi.org/10.3758/bf03197238.

López, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology, 32*, 251–295. http://dx.doi.org/10.1006/cogp.1997.0651.

López, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category-based induction. *Child Development, 63*, 1070–1090. http://dx.doi.org/10.2307/1131519.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 10*, 317–332. http://dx.doi.org/10.3758/bf03196515.

Medin, D. L., Waxman, S., Woodring, J., & Washinawatok, K. (2010). Human-centeredness is not a universal feature of young children's reasoning: Culture and experience matter when reasoning about biological entities. *Cognitive Development, 25*, 197–207. http://dx.doi.org/10.1016/j.cogdev.2010.02.001.

Mervis, C., & Pani, J. (1980). Acquisition of basic object categories. *Cognitive Psychology, 12*, 496–522. http://dx.doi.org/10.1016/0010-0285(80)90018-3.

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: MIT Press.

Nagel, E. (1939). *Principles of the theory of probability.* Chicago: University of Chicago Press.

Osherson, D. N., Smith, E. E., Wilkie, O., López, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185–200. http://dx.doi.org/10.1037/0033-295x.97.2.185.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 811–828. http://dx.doi.org/10.1037//0278-7393.26.4.811.

Rhodes, M. (2012). When children ignore evidence in category based induction: Irrational inferences? *Advances in Child Development and Behavior, 43*, 219–235.

Rhodes, M., & Brickman, D. (2010). The role of within-category variability in category-based induction: A developmental study. *Cognitive Science, 34*, 1561–1573. http://dx.doi.org/10.1111/j.1551-6709.2010.01137.x.

Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition, 108*, 543–556. http://dx.doi.org/10.1016/j.cognition.2008.03.002.

Rhodes, M., & Gelman, S. A. (2009). Five-year-olds' beliefs about the discreteness of category boundaries for animals and artifacts. *Psychonomic Bulletin & Review, 16*, 920–924. http://dx.doi.org/10.3758/pbr.16.5.920.

Rhodes, M., Gelman, S. A., & Brickman, D. (2008). Developmental changes in the consideration of sample diversity in inductive reasoning. *Journal of Cognition and Development, 9*, 112–143. http://dx.doi.org/10.1080/15248370701836626.

Rhodes, M., Gelman, S. A., & Brickman, D. (2010). Children's attention to sample composition in learning, teaching and discovery. *Developmental Science, 13*, 421–429. http://dx.doi.org/10.1111/j.1467-7687.2009.00896.x.

Rhodes, M., Gelman, S. A., & Karuza, J. C. (2014). Preschool ontology: The role of beliefs about category boundaries in early categorization. *Journal of Cognition and Development, 15*, 78–93. http://dx.doi.org/10.1080/15248372.2012.713875.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14*, 665–681. http://dx.doi.org/10.1016/s0022-5371(75)80055-7.

Rosner, S., & Hayes, D. (1977). A developmental study of category item production. *Child Development, 48*, 1062–1065. http://dx.doi.org/10.2307/1128361.

Ross, N., Medin, D., Coley, J., & Atran, S. (2003). Cultural and experiential differences in the development of folkbiological induction. *Cognitive Development, 18*, 25–47. http://dx.doi.org/10.1016/s0885-2014(02)00142-9.

Shipley, E. F., & Shepperson, B. (2006). Test sample selection by preschool children: Honoring diversity. *Memory & Cognition, 34*, 1444–1451. http://dx.doi.org/10.3758/bf03195909.

Shtulman, A., & Schulz, L. (2008). The relation between essentialist beliefs and evolutionary reasoning. *Cognitive Science, 32*, 1049–1062. http://dx.doi.org/10.1080/03640210801897864.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America, 105*, 5012–5015.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*, 245–272. http://dx.doi.org/10.1037/0033-295x.114.2.245.

Zhong, L., Lee, M., Huang, Y., & Mo, L. (2014). Diversity effect in category-based inductive reasoning of young children: Evidence from two methods. *Psychological Reports, 114*, 198–215. http://dx.doi.org/10.2466/10.11.pr0.114k17w3.