Continuity and change in the development of category-based induction

Marjorie Rhodes\textsuperscript{a*} & Peter Liebenson\textsuperscript{b}

\textsuperscript{a}New York University, 6 Washington Place, 301, New York, NY 10003;
marjorie.rhodes@nyu.edu; 212-998-3546

\textsuperscript{b}New York University, 6 Washington Place, 302, New York, NY 10003;
peter.liebenson@nyu.edu

*Corresponding author
The present research examined the extent to which the mechanisms that underlie induction stay constant across development or undergo fundamental change, by testing how children (ages 5-10) determine the informative value of samples of evidence. How children evaluated evidence varied across age and type of category. For familiar natural kinds, children did not view diverse evidence as more informative than non-diverse evidence until ages 9-10. In contrast, for novel categories, children of each age made stronger 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, and suggest that developmental changes in induction stem from developmental differences in the nature of conceptual structures.

**Key words:** Inductive reasoning, conceptual development, diversity-based reasoning, category-based induction, conceptual structure
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 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). The development of category-based induction is central to our understanding of cognitive development—particularly whether development is best understood as involving fundamental conceptual change, changes in learning mechanisms, or simply the elaboration and accumulation of knowledge (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 regarding the entire kind (e.g., mammals), whereas less diverse samples (e.g., a tiger and a lion) are informative regarding more limited subsets. Diverse samples are more informative than less diverse samples because they provide broader 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 view diverse samples as more informative than non-diverse ones (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 find out if there is support for category-wide generalizations, preferring to assemble samples composed of a lion and a hamster, for example, rather than a lion and a tiger, to discover properties of mammals (Kim & Keil, 2003; López, 1995; López, Atran, Coley, Medin, & Smith, 1997; Rhodes, Gelman, & Brickman, 2008; Rhodes, Brickman, & Gelman, 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 variability 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; Heit & Feeney, 2007; Lopez 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 childhood, with the aim of identifying what accounts for developmental changes that have been reported in prior work. We aim to clarify whether age-related changes found in prior work reflect
real fundamental developmental changes in the processes that support induction. In contrast to adults, children below the age of 9 often view diverse and non-diverse samples as equivalently informative 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; López, Gelman, Gutheil, & Smith, 1992; Li, Cau, Li, Li, & Déák, 2009). 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 et al., 2008; Rhodes, Gelman, & Brickman, 2010; Zhong, Lee, Huang, & Mo, 2014). In these studies, diversity-based reasoning has been found to emerge around ages 8-9 and become more robust by ages 10-11.

Several accounts have been proposed for why younger children fail to consider sample diversity and what these findings mean for the nature of cognitive development. Broadly, computational capacity hypotheses suggest that children’s neglect of sample diversity stems from underlying developmental differences in computational abilities. For example, younger children might have difficulty generating the relevant inclusive categories from particular samples (Lopez 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 should not engage in diversity-based reasoning under any circumstances, as they simply
lack the required cognitive skills to incorporate diversity into their sample evaluations. If this approach is correct, then the mechanisms that underlie knowledge acquisition change fundamentally across development.

Alternately, *continuity hypotheses* suggest that children at least as young as four or five can successfully consider sample diversity in a normative and adult-like manner, as long as the experimental tasks are manageable for them (Heit & Hahn, 2001; Shipley & Shepperson, 2006; Zhong et al., 2014). From this perspective, younger children have the underlying abilities to appropriately consider sample diversity, but sometimes fail to do so either because experimental tasks are too challenging (e.g., in terms of information processing demands; Heit & Hahn, 2001) or because they lack some relevant background knowledge (e.g., they might lack knowledge of taxonomic hierarchies which makes it difficult to compute how well a sample represents a broader kind; Carey, 1985). From this perspective, the mechanisms that underlie induction are continuous across development, but children may make more faulty inferences because they cannot access these mechanisms as readily.

In support of continuity hypotheses, there are some positive reports of diversity-based reasoning in younger children using simplified tasks (Heit & Hahn, 2001; Shippley & Shepperson, 2006, Zhong et al., 2014). Careful consideration of these experimental tasks, however, 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 young children, see Rhodes, Brickman et al., 2008; Rhodes, Gelman et al. 2008).

Inconsistent with continuity hypotheses, children fail to consider sample diversity even when they can clearly manage the experimental tasks. For example, Rhodes, Brickman et al. (2008) showed that 6-year-olds reliable sought out typical evidence, but in the same paradigm, failed to seek out diverse evidence (also shown by Zhong et al., 2014). Thus, children could understand the experimental tasks and respond systematically, but nevertheless failed to consider sample diversity. Similarly, Rhodes, Gelman et al. (2008) showed that children could reliably distinguish diverse from non-diverse samples and make judgments based on this information but did not prefer to test diverse samples to draw inferences about a broader kind.

Although continuity hypotheses cannot account for the full scope of children’s failures in prior work, there is also reason to suspect that computational capacity approaches might not be able to explain all of children’s successes. As reviewed earlier, computational capacity accounts predict that young children lack the requisite cognitive skills to appreciate the value of diverse samples. Yet, Rhodes et al. (2010) found that 5-year-olds, in identical paradigms, considered sample diversity to draw an inference about a communicator’s intent (see also Xu & Tenenbaum, 2007), but not 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, Lopez et al., 1992; Guthiel & Gelman, 1997). 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 dalmations), children used sample
diversity in an adult-like manner (inferring that the property applied only to dalmations 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 a computational capacity account could explain children’s varying success and failure in this paradigm.

Here we test a new hypothesis regarding the development of induction, aiming to explain both children’s successes and failures. We propose that developmental changes in induction stem from developmental differences in the nature of conceptual structure. In particular, we propose that younger children’s representations of certain types of categories are more uniform than those held by older children and adults, thus leading them to neglect the value of samples that provide diverse representation of broader kinds. In early childhood, such a uniformity bias could help get knowledge acquisition off the ground, by allowing children to focus on simpler, easier to learn concepts (Berndt & Heller, 1986; Gelman, 1988, 2003; Gelman & Kalish, 1993; Kalish, 1998; Rhodes & Gelman, 2008, 2009a; Taylor, 1996; Taylor, Rhodes, & Gelman, 2009). Yet, a uniformity bias would also lead children to overlook meaningful and important variation and thus to neglect the importance of collecting samples that provide coverage of that variation.

In support of this hypothesis, inducing 7-year-old children to focus on within-category variation increased their tendency to seek out diverse evidence (Rhodes & Brickman, 2010). Yet, this prior work leaves a number of issues unaddressed. First, Rhodes and Brickman (2010) tested how children sought out evidence, not their actual inductive inferences (how children use sample diversity to determine a proper scope of generalization). Second, this work involved 7-year-old children, who are relatively close to the age at which children have demonstrated diversity-based reasoning in previous reports (Li et al., 2009; López et al., 1992; Rhodes, Brickman et al., 2008).
Thus, it remains unknown whether younger children are constrained by more entrenched cognitive limitations.

The present research compares the development of diversity-based reasoning across different types of categories. Much of the research on the development of induction has focused on familiar biological kinds. These categories have been the subject of interest for many reasons—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). If a uniformity bias interferes with diversity-based reasoning, however, then such a bias might be particularly strong for biological categories, as young children have strong expectations that these categories mark coherent, homogeneous natural kinds (Gelman, 1988; Rhodes & Gelman, 2009a, 2009b). Thus, consideration of sample diversity might emerge earlier for other types of categories. The present work compares diversity-based reasoning for categories with different ontological statuses (fictional vs. commonplace; animal vs. artifact). Thus, this work tests how the mechanisms that support induction vary across both development and domain.

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, fictional animals and artifacts. Study 3 directly compares children’s reasoning about a familiar animal category and a novel, fictional category using identical perceptual stimuli. These studies include children ages 5-10, spanning the ages at which diversity-based reasoning has been found to develop. If children’s difficulties with diversity-based reasoning stem from underlying computational limitations or depend only on task demands, then we should find similar patterns across these three studies. Alternately, if
domain-specific conceptual structure interferes with diversity-based reasoning in early childhood, then performance—particularly of younger children—should vary by the type of category.

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 the informative value of two samples, 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 Figure 1). To account for any biases that 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.

Figure 1. Sample trial, Study 1.
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. I’m going to tell you a bunch of special things about birds. 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 Figure 2).

**Figure 2.** Collages to illustrate target categories, Studies 1-3.

For each trial (see Figure 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. 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 canary). Items were blocked by sample-type, and 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 viewed 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 (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.

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

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

* For study 3, this was replaced with, "have a special thing inside called a furcula."

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 unrelated 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 people 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 and age entered as a continuous covariate. This analysis yielded a main effect of sample (diverse vs. non-diverse), $F(1, 69) = 5.65, p = .02$, and a sample*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.

To investigate the nature of the age*sample interaction, 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; $M =$
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 Figure 3.

**Figure 3.** Children’s generalizations to birds from diverse and non-diverse samples, by age, Study 1.

![Figure 3](image)

*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. In particular, children of all ages generalized more to birds than to frogs. For each age group, we conducted 2(sample: non-diverse, diverse) X 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 Figure 4.
Figure 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.

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 generalization 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?” 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 $X^2 (1) = 73.16, p < .001$; ages 7-8, Wald $X^2 (1) = 35.98, p < .001$; ages 9-10, Wald $X^2 (1) = 32.51, p < .001$). 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, $p s < .001$. 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 view these samples as more informative.

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>5 - 6 year olds</th>
<th>7 - 8 year olds</th>
<th>9 - 10 year olds</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diverse samples</td>
<td>0.94 (0.67 - 0.97)</td>
<td>0.77 (0.60 - 0.88)</td>
<td>0.84 (0.48 - 0.77)</td>
</tr>
<tr>
<td>Non-diverse samples</td>
<td>0.11 (0.05 - 0.24)</td>
<td>0.04 (0.01 - 0.11)</td>
<td>0.04 (0.01 - 0.12)</td>
</tr>
<tr>
<td><strong>Study 2 (animal condition)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diverse samples</td>
<td>0.83 (0.61 - 0.94)</td>
<td>0.75 (0.58 - 0.87)</td>
<td>0.78 (0.57 - 0.90)</td>
</tr>
<tr>
<td>Non-diverse samples</td>
<td>0.11 (0.04 - 0.29)</td>
<td>0.07 (0.02 - 0.19)</td>
<td>0.02 (0.00 - 0.14)</td>
</tr>
<tr>
<td><strong>Study 2 (artifact condition)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diverse samples</td>
<td>0.75 (0.58 - 0.87)</td>
<td>0.73 (0.56 - 0.85)</td>
<td>0.73 (0.54 - 0.86)</td>
</tr>
<tr>
<td>Non-diverse samples</td>
<td>0.05 (0.01 - 0.19)</td>
<td>0.03 (0.00 - 0.18)</td>
<td>0.09 (0.03 - 0.24)</td>
</tr>
<tr>
<td><strong>Study 3 (bird condition)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diverse samples</td>
<td>0.87 (0.69 - 0.95)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-diverse samples</td>
<td>0.33 (0.18 - 0.53)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Study 3 (model condition)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Diverse samples</td>
<td>0.70 (0.55 - 0.82)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-diverse samples</td>
<td>0.19 (0.09 - 0.33)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 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 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.

| 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.

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 \( X^2(1) = 13.84, p < .001 \)).

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, showing an adult-like pattern. 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, 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 and 3.

3.1. STUDY 2

To test if domain-specific conceptual structures interfere with diversity-based reasoning among children, Study 2 examined how the processes that support induction vary by the ontological status of the target categories. In Study 2, all children were introduced to a fictional category, modies, which by random assignment were described as animals or artifacts.

Study 1 asked children about a familiar natural kind, for which they are likely to already know about many shared properties, and for which they have likely been exposed to extensive generic language (e.g., “birds fly”; Gelman, Coley, Rosengren, Hartman, & Pappas, 1998; Gelman, Goetz, Sarnecka, & Flukes, 2008). Generic language leads children to assume that category members are similar in known and yet-to-be-discovered ways (Gelman, Ware, & Kleinbern, 2010; Rhodes, Leslie, & Tworek, 2012) and to exclude variation from concepts (Gelman & Raman, 2003). Thus, a uniformity bias might operate particularly strongly for familiar natural kinds. If such a bias interferes with diversity-based reasoning, children should be
more likely to engage in diversity-based reasoning for a novel category, about which they have no prior knowledge, have not been extensively exposed to generic language, and thus have weaker expectations regarding category structure (Gelman et al., 2010). Further, because children expect animal categories to be more homogeneous than artifact categories even when first learning about fictionalized kinds (Brandone & Gelman, 2009; Rhodes, Gelman, & Karuza, 2014), they may be particularly likely to engage in diversity-based reasoning for fictionalized artifacts. Alternately, if younger children are either limited by computational capacities or impaired by task demands, then we should find the same pattern of development as in Study 1 (as a task of identical structure is used here).

3.2. Methods

3.2.1. Participants

Study 2 included 105 children (70 female, 35 male). 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. Children were drawn from 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). Participants at each age group were randomly assigned to hear modies described as animals or artifacts.

3.2.2. Scale Training and Category Introduction

First, children completed the same training on the dot scale as in Study 1 (see Figure 1). Then they were introduced to the novel category, modies. To do so, children were shown three modies (see Figure 5), which were perceptually identical across condition but were described as either artifacts or animals (see Brandone & Gelman, 2009; Rhodes et al., 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.”

**Figure 5.** Stimuli used to introduce the target categories, Studies 2 and 3.

![Figure 5](image)

**3.2.3. Induction Task**

Next, a collage showing many modies was brought out (see Figure 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.”
Subsequently, children completed induction trials, following an identical structure to those in Study 1 (see Figure 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).

3.2.4. 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).

3.2.5. 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.

3.3. 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.

3.3.1. 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 (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 $D = .44$. Planned comparisons confirmed that the effect of sample held up at each age (see Figure 6). Follow-up analyses confirmed no interactions between category-type and sample in any age group examined separately.

Figure 6. Children’s generalizations from diverse and non-diverse samples, by age, Study 2.
3.3.2. 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, $t(29) = 11.16, p < .001$).

3.3.3. 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) X 2 (sample) X 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.

3.4 Discussion

In contrast to Study 1, Study 2 revealed diversity-based reasoning among children aged 5-8 (as well as aged 9-10, as in Study 1). This is the first study to provide clear evidence of diversity-based reasoning in young children—in particular, that children view diverse samples as more informative than non-diverse samples for drawing inferences about broader kinds. 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 fictional category for which they held no prior knowledge. Thus, children’s concepts of familiar biological kinds appear to interfere with diversity-based reasoning.

4.1. STUDY 3

The difference in children’s performance across Studies 1 and 2 suggests that children’s beliefs about familiar biological kinds lead them to neglect sample diversity. However, 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. Additionally, because Study 2 presented a constructed category, perhaps the stimuli in Study 2 were more transparent regarding how the diverse sample reflected the variability in the broader set. Thus, the aim of Study 3 was to compare children’s reasoning about a familiar natural kind (birds) and a novel fictional kind (modies) using identical
perceptual stimuli and procedures. 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.

4.2. Methods

Participants included 36 children ages 5.0-6.9 (M age = 6.0, 18 male, 18 female). Of the 36 participating children, 30 parents filled out the supplementary background questionnaire used in Study 2. These data showed that 17 participants lived in urban environments, 12 in suburban, and 1 in rural; 20 participants were from New York, 4 from New Jersey, 5 from other states, and 1 from Canada. 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.

4.2.1. Scale Training and Category Introduction

Participants were randomly assigned to bird or modie conditions. 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 Figure 1). The critical exception was in the introduction to the category. In both conditions, children were shown identical perceptual stimuli (see Figure 3).

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 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 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).

4.2.2. 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 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 modies fell down and got a little scrape last week.” “These three 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?”

4.3. Results and Discussion

We ran a 2 (sample) X 2 (category-type) repeated measures analysis of variance on children’s generalizations following diverse and non-diverse samples. To test for possible developmental changes across this age range, we included exact age as a continuous covariate. These analyses revealed a sample X category-type interaction, $F(1, 33) = 4.53, p = .04$ (see Figure 7), with no main or interactive effects of age. As shown in Figure 7, children generalized more from diverse samples than non-diverse samples (Cohen’s D = .31) only when stimuli were referred to as “modies.” 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 natural kind.
Figure 7. 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, one tailed (to test a directional hypothesis, given the effect found in Study 2).

As in previous studies, children were more likely to respond that the diverse samples ($M = .80$, $CI = .68-.88$) than the non-diverse samples ($M = .25$, $CI = .16-.37$) appeared “pretty different from each other,” Wald $X^2 (1) = 51.00$, $p < .001$. 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 ($ps > .60$). Overall children tended not to generalize the incidental properties to many other category members ($M = 1.07$, $SE = .15$). Confirming that children understood the conditions as intended, when asked to identify the stimuli, all 18 children in the bird condition responded with “birds,” whereas in the modie condition, none of the children identified the stimuli with birds. Responses varied; for
example, children responded with: “a kind of magical creature,” “a stick that’s alive,” “like a worm.”

5.1. 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 that underlie this component of induction vary across development and ontological domain. 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. Only at ages 9-10 do children generalize more from diverse samples. Yet, for unfamiliar fictional animals and artifacts, younger children (ages 5-8) successfully engaged in diversity-based reasoning. Study 3, which used identical perceptual stimuli to depict a familiar animal category and a fictitious one, confirms that the youngest children’s differential responding to familiar and unfamiliar kinds stems from the conceptual knowledge they bring to the task, not domain differences in the perceptual properties of the stimuli.

Children’s ratings of the diversity of the stimuli show that simple perception of diversity does not explain when children do or do not incorporate diversity into their sample evaluations. Children of each age and in every condition rated the diverse sets as more diverse than the non-diverse sets. Further, older children in Study 1—the only participants to successfully engage in diversity-based reasoning in that study—actually rated the diverse sets of birds as less perceptually diverse than did the youngest children. Across studies, children did not rate the diverse modies in Study 2 as more diverse than the diverse birds in Study 1, and they also did not rate the diverse stimuli as more diverse in the modie condition than in the bird condition of Study
3. Thus, the extent to which younger children perceived diversity does not explain the pattern of findings across conditions or studies.

The present Studies 2-3 are the first to clearly demonstrate that children aged 5-8 can engage in diversity-based reasoning. Whereas previous work has shown that children of these ages can distinguish diverse from non-diverse samples and sort evidence based on diversity (Heit & Hahn, 2001; Zhong et al., 2014), Studies 2 and 3 are the first to clearly reveal that children aged 5-8 can generalize further from diverse samples 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. These findings are inconsistent with computational capacity approaches to explaining children’s difficulties with diversity-based reasoning. Because identical tasks were used to assess children’s reasoning about familiar and fictitious kinds (see especially Study 3), children’s differential performance across the different studies and conditions in this research cannot be explained by differences in computational demands.

Yet because performance varied across conditions that used identical tasks, these results also show that children’s difficulty with diversity-based reasoning does not stem from task demands. Younger children in Study 1 and in the bird condition of Study 3 failed to view diverse samples as more informative than non-diverse samples, even though their successful performance in Study 2 and the modie condition of Study 3 (along with reasonable responses to all control questions) reveals that they could cope with the task demands and display diversity-based reasoning under these circumstances.

The present studies are not only the first to show successful diversity-based reasoning in younger children; they also provide a framework for understanding when children succeed and
fail on these tasks. Children below the age of nine appear not to consider sample diversity for familiar biological kinds, but do so for novel animal and artifact categories. We have proposed that these differences stem from developmental and domain differences in the extent to which a uniformity bias shapes children’s concepts. There are two mechanisms by which such a uniformity bias could operate, which will be important to distinguish in future work.

First, a uniformity bias could operate via a tendency to expect certain categories to be highly homogeneous (i.e., to assume that all birds are fundamentally similar to each other). On this account, children recognize atypical instances of categories (e.g., ostriches), but expect these atypical inferences to be fundamentally similar to other category members. This could explain why children would not recognize the informative value of samples that provide diverse representation of categories—if all birds are viewed as fundamentally alike, then all samples could be considered equivalently informative. This explanation—and the present finding that younger children fail to engage in diversity-based reasoning particularly for familiar natural kinds—is consistent with prior work showing that young children expect familiar natural kinds (but not other types of categories) to be more homogeneous than older children or adults do (Gelman, 1988). For example, when asked questions about familiar natural kinds, 5-year-olds expect all category members to display category-typical properties, even when given contrasting information about particular individuals (Berndt & Heller, 1986; Taylor, 1996; Taylor et al., 2009). Children hold homogeneity expectations more strongly for familiar natural kinds than for either fictional animals (Gelman et al. 2010) or for real or fictional artifacts (Brandone & Gelman, 2009; Rhodes & Gelman, 2009a, 2009b; Rhodes et al., 2014). From this perspective, across development, children incorporate more within-category variation into their biological concepts.
There are several features of the present data that are not entirely consistent with this account. First, if younger children viewed birds as more homogeneous than older children did, we might expect them to generalize more overall. This was not the case—in fact, overall, older children in Study 1 generalized to more birds than did younger children (see Figure 4 and Table A1). Second, even for fictitious categories, children expect animal categories to be more homogeneous than artifact categories (Brandone & Gelman, 2009; Rhodes et al., 2014). Thus, if expectations of homogeneity interfered with diversity-based reasoning, we might have expected to find more diversity-based reasoning for the fictitious artifacts than animals, but this also was not the case. Nevertheless, it would be useful in future research to test the extent to which expectations of homogeneity influence diversity-based reasoning directly, by experimentally inducing such expectations for fictitious kinds, and testing whether inducing these beliefs interferes with diversity-based reasoning.

A second way that a uniformity bias might operate is via category structure itself, such that children’s concepts consist of a more uniform set of exemplars than adults’ concepts do. If younger children’s concepts include fewer atypical instances, their prototypes would shift towards relatively more extreme values. For example, if children exclude the instance “ostrich” from the category “bird,” their estimation of the typical size of a bird would be shifted towards smaller values than if they included this atypical instance. Children could also focus on extreme values because they are useful for distinguishing categories from one another (Davis & Love, 2010); for example, accentuating the average difference in size between birds and mammals could facilitate category learning—an important task of early childhood. Further, children’s essentialist beliefs—which can include expectations that natural categories reflect ideal forms (Gelman & Rhodes, 2012)—could lead them to focus on category members that approximate
these ideal forms (e.g., birds with the best, instead of average, flying ability). According to Osherson et al. (1990) diversity and typicality influence induction in a similar manner, by increasing the coverage provided by a particular sample. If children’s prototypes are shifted towards extreme values, however, then children may view samples that are closer to their prototypes as more informative than those that provide diverse coverage of the adult concept.

Consistent with this interpretation, typicality structures—particularly the extent to which they are predicted by measures of central tendency or category ideals—vary by levels of knowledge, expertise, and the type of category (Barsalou, 1985; Burnett, Medin, Ross, & Blok, 2005; Johnson, 2001; Kim & Murphy, 2011; Lynch, Coley, & Medin, 2000; Proffitt et al., 2000; Voorspoels, Vanpaemel, & Storms, 2011; Voorspoels, Storms & Vanpaemel, 2013). Also—unlike diversity-based reasoning—typicality effects on category learning and induction have been well documented in early childhood (Bjorklund & Thompson, 1983a, 1983b; Southgate & Meints, 2000). By at least age five, children generalize further from typical samples than atypical ones for familiar natural kinds (López et al., 1992) and prefer to seek typical evidence to test their category-based inferences (e.g., preferring to test a golden retriever over a Chihuahua to learn about dogs; Rhodes, Brickman et al., 2008; Zhong et al., 2014).

From this perspective, children fail to engage in diversity-based reasoning for familiar natural kinds because the structure of their concepts renders it an uninformative strategy. In contrast, when their concepts are not centered on strong prototypes (such as unfamiliar categories for which they do not know the typicality structure or in domains where they do not have essentialist beliefs that emphasize ideal forms) they instead attend to samples that provide broad coverage. These possibilities could be tested in future work by creating a typicality structure of
the novel category and testing how this affects children’s responses, as well as by testing an even broader array of categories (e.g., familiar artifacts).

Identifying which of the above processes accounts for children’s failure to consider sample diversity for familiar natural kinds will speak to the extent to which the processes that underlie induction depend on developmental factors, ontological domain, and/or on one’s stage of concept learning. In the studies reported here, when conceptual knowledge was equated across age—by using fictitious categories in Study 2—we found no developmental changes in diversity-based reasoning. Thus, if children of all ages were taught more about the categories—either via generic language to increase expectations of homogeneity or creation of a typicality structure—perhaps children of each age (and even adults) would pass through a stage of concept learning where they neglect sample diversity in order to make rapid inferences based on typical or ideal category members. Later, with increased knowledge and familiarity with the category, perhaps participants would return to considering sample diversity. Alternately, there could be additional developmental factors at play (e.g., tendencies for younger children to consider only one strategy at a time, or developmental differences in category structure) that contribute particularly to younger children’s performance.

The present work was 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 stages of concept learning or is constrained by other developmental factors (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 learning show important continuity. Yet, these studies also revealed important developmental change in when children access these mechanisms. Younger children did so for fictitious categories, but not for familiar natural kinds. This research suggests that developmental changes in induction stem from developmental differences in the nature of conceptual structure. Identifying precisely which features of conceptual structure explain these phenomena remains an important topic for future work.
Notes

1To 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.

2All 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.
Acknowledgments

This research was supported by National Science Foundation grant BCS-1147543. We thank Samantha Bordoff, Annie Chen, Sarah Cordell, Zach Genduso, Nicole Graziano, Emily Foster Hanson, and Christina Tworek for assistance with data collection, Danny Zeiger and the staff of the Discovery Room at the American Museum of Natural History for their support of this research, and Dr. Gregory Murphy for comments on a previous draft.
References


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


cues to interpret generics. *Child Development, 74*, 308-325. doi:10.1111/1467-8624.00537

essentialism impedes evolutionary understanding. In K.R. Rosengren, S. Brem, E.M.
Evans, & G. Sinartra (Eds). *Evolution challenges: Integrating research and practice in
teaching and learning about evolution* (pp. 3-21). New York: Oxford University Press.
doi:10.1093/acprof:oso/9780199730421.003.0001

doi:10.1080/15475440701542625

Gelman, S.A., Ware, E.A., & Kleinberg, F. (2010). Effects of generic language on category
doi: 10.1016/j.cogpsych.2010.06.001

doi:10.1006/jecp.1996.2344

*Inductive Reasoning: Experimental, Developmental and Computational Approaches* (pp.
25-54). Cambridge: Cambridge University Press. doi:10.1017/cbo9780511619304.003

doi:10.3758/bf03212996


Appendix

Table A1

<table>
<thead>
<tr>
<th>Study</th>
<th>5 - 6 year olds</th>
<th>7 - 8 year olds</th>
<th>9 - 10 year olds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>1.73 (0.20)</td>
<td>1.73 (.21)</td>
<td>2.32 (.21)</td>
</tr>
<tr>
<td>Study 2</td>
<td>1.89 (0.14)</td>
<td>1.67 (0.12)</td>
<td>1.51 (0.14)</td>
</tr>
<tr>
<td>Study 3 (bird condition)</td>
<td>1.70 (0.22)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 3 (modie condition)</td>
<td>2.13 (0.22)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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