CHILD DEVELOPMENT

Child Development, July/August 2016, Volume 87, Number 4, Pages 1090–1098

Children Prefer Diverse Samples for Inductive Reasoning in the Social Domain

Alexander Noyes and Stella Christie
Swarthmore College

Not all samples of evidence are equally conclusive: Diverse evidence is more representative than narrow evidence. Prior research showed that children did not use sample diversity in evidence selection tasks, indiscriminately choosing diverse or narrow sets (tiger–mouse; tiger–lion) to learn about animals. This failure is not due to a general deficit of inductive reasoning, but reflects children's belief about the category and property at test. Five- to 7 year-olds' inductive reasoning (n = 65) was tested in two categories (animal, people) and properties (toy preference, biological property). As stated earlier, children ignored diverse evidence when learning about animals' biological properties. When learning about people's toy preferences, however, children selected the diverse samples, providing the most compelling evidence to date of spontaneous selection of diverse evidence.

Inductive reasoning requires the ability to discriminate between better and worse samples of evidence—a diverse sample, for instance, is more representative than a narrow one, thus more useful for drawing broader generalizations (Heit, 2000; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). To find out if all animals in a zoo have contracted a new disease, it is more informative to examine the disease's presence in a diverse sample of animals (e.g., a lion and a mouse) than in a narrow sample (e.g., a lion and a tiger). The ability to use sample diversity to evaluate an inductive inference about a broader kind is a standard measure of children's mature inductive reasoning. A typical task tests children's evaluation of samples by asking them to decide from which sample to extend a property. For example, in López, Gelman, Gutheil, and Smith (1992), children learned that a diverse sample (a cat and a buffalo) had property X, whereas the narrow sample (a cow and a buffalo) had property Y. They had to decide if all animals had property X or Y. Five- and six-year-olds did not find the diverse sample more compelling and extended properties X and Y to all animals equally often. Indeed, 6-year-old children's lack of preference for diverse samples in inductive reasoning tasks about broader kinds has been robustly shown in many studies (Gutheil & Gelman, 1997; Li, Cao, Li, Li, & Deak, 2009; Rhodes & Brickman, 2010; Rhodes, Brickman, & Gelman, 2008; Rhodes, Gelman, & Brickman, 2008). These studies have included various methodological approaches, for example, varying location diversity (e.g., a sample from one mountain vs. a sample from four different mountains) and asking children to explicitly select the more informative sample. Across methodological approaches, however, 6-year-old children select between diverse and narrow evidence at chance, showing no preference for diverse samples of evidence. It is not until 9 years of age that children demonstrate a spontaneous preference for diverse evidence (López et al., 1992).

Broadly speaking, there are two possible explanations for why younger children do not select diverse evidence. First, 6-year-olds may have a deficit in their reasoning capabilities—they do not yet grasp the representativeness of evidence (Li et al., 2009; López et al., 1992; Rhodes, Brickman, et al., 2008). Young children may lack the concept of "coverage" that some samples of evidence (e.g., lion/mouse) cover more of the category space than other samples (e.g., lion/tiger). This calculation is essential for discriminating between diverse and

This research was supported by a Swarthmore College Faculty Research grant and a Lang Sabbatical Fellowship to Stella Christie and by a summer research fellowship from the Howard Hughes Medical Institute to Alexander Noyes. We thank the children, families, schools, and organizations for their participations in this research. We also thank Abigail Robinson and members of the Swarthmore College Cognition and Development Laboratory for their assistance in recruiting participants.

Correspondence concerning this article should be addressed to Stella Christie, Department of Psychology, Swarthmore College, 500 College Ave., Swarthmore, PA 19081. Electronic mail may be sent to christie@swarthmore.edu.

 0009-3920/2016/8704-0011
DOI: 10.1111/cdev.12522

narrow samples of evidence (Osherson et al., 1990). Without it, children would not be able to do diversity-based reasoning regardless of the category or property in question. This explanation posits that 6-year-old children's failure to use diversity-based reasoning is context independent; they will not prefer diverse evidence in any context.

In contrast, the second explanation suggests that children's reasoning is intact but that they reason from a set of different beliefs than adults do (Gutheil & Gelman, 1997; Heit & Hahn, 2001; Rhodes & Brickman, 2010; Rhodes, Brickman, et al., 2008). Adults believe that the lion/mouse sample is more representative than the lion/tiger sample because lions and mice share fewer common properties than lions and tigers do. Analogous to a Venn diagram calculation, a smaller area of intersection (fewer shared properties) implies a larger overall union (larger coverage)—hence, a more representative sample. Children may be able to do this calculation, but unlike adults, they believe that lions are just as likely to share properties with mice as with tigers. Under this belief, there is no difference of coverage between the diverse and narrow sets, and thus children choose at chance. Importantly, the second explanation suggests that children's inductive reasoning in evidence selection tasks should be context dependent. Under the appropriate context, even 6-year-olds should be able to employ diversity-based reasoning. Indeed, there is some support for this explanation: Rhodes and Brickman (2010) found that 7-year-old children could implement diversity reasoning when first primed to focus on within-category heterogeneity. When children were guided to think about the many differences that exist within a single category (e.g., birds differ on their color, size, and ability to fly), they showed preference for the diverse samples of evidence. Rhodes and Brickman argue that without the prime, children either do not know how heterogeneous animals are or have strong assumptions that animals are homogenous. With such beliefs about animal homogeneity, children reasonably do not incorporate diversity into their decisions.

If the second explanation is correct—that younger children fail because of their beliefs about animals and not because of a reasoning deficit—then they should be able to spontaneously select diverse evidence in the right context (i.e., when they believe samples differ on the likelihood of sharing the property). We hypothesize that people's toy preferences may be one such context. Social categories like age, race, and gender are highly salient for even young children and provide a basis for decision making as early as 2–3 years for gender and 4–5 years for race (Kinzler, Shutts, & Correll, 2010). Young children also use gender, age, and race as a way of predicting their own and others' toy preferences (Martin, Eisenbud, & Rose, 1995; Shutts, Banaji, & Spelke, 2010). This body of research shows that young children believe that two individuals are more likely to share toy preferences when they are both of the same age, race, or gender. That is, unlike the homogeneity beliefs with animals, children believe that two same-gender people have more things in common than different-gender people. If children employ diversity-based reasoning under such a belief, they should spontaneously choose the diverse (e.g., Black and White children) over the narrow sample of evidence (e.g., two White children) when asked to generalize to all children.

This argument is similar to Heit and Hahn (2001) and Shipley and Shepperson (2006) who argued that children implement diversity-based reasoning in the artifact domain. However, these studies have been criticized on methodological grounds (Rhodes, Gelman, et al., 2008). In Heit and Hahn, children were asked to decide whether a novel toy (e.g., basketball) belonged to a child who owned only one kind of type (e.g., many baseballs) or a child who owned many types of toy (e.g., tennis ball, baseball, soccer ball). Children's success in this study could be accounted by them simply using a matching strategy (e.g., someone who only likes baseballs would not like basketballs) rather than a generalization strategy. We addressed this problem by asking children to evaluate the evidence itself (asking which of the two samples is more informative) rather than asking them to generalize to a new exemplar. This way children cannot simply match the new exemplar to one of the two samples. In Shipley and Shepperson, children tested a diverse set (e.g., blue and red whistles) to determine if a party favor (whistles) worked properly. But children may have merely tested the two subpopulations (red and blue whistles) rather than thinking about the generalizability of the property. We addressed this concern by explicitly asking children to generalize to a broader category (all children). Children could not simply "check" the subpopulations because no single trial included all the subpopulations represented in the study (e.g., a trial of Black and White boys excludes Black and White girls).

In sum, we hypothesized that children's ability to evaluate the quality of evidence along the axis of diversity depends on their beliefs about the distribution of properties within the category. To test this, children were presented with samples of animals (animal) or samples of people (people). In both animal and people conditions, children selected evidence about toy preferences (toy) and internal biological properties (hormones). They had to select the best sample of evidence to infer if all people or animals possessed a given property. This is a rigorous task, because it requires children to explicitly acknowledge which sample of evidence—diverse versus narrow—is more informative. As in past literature, we anticipated that children would not demonstrate a preference for diverse samples when reasoning about animals' internal properties. Since we do no expect children to believe that toy preferences vary more among diverse animals or that unknown biological hormones vary more among diverse people, we do not predict children to choose diverse over narrow evidence in these two contexts. But crucially, we predicted that children would prefer diverse evidence when reasoning about people's toy preferences. This success hinges specifically on children's belief about a property's distribution within a category.

Method

Participants

Sixty-five 5- to 7-year-olds (40 male, 25 female; $M = 6.09$ years, $SD = 0.840$, range = 5.0–7.9 years) participated in the study. One child is excluded from the final analysis because of experimenter error. Children were recruited from local preschools and kindergartens in a Northeastern college town between June 2013 and May 2014; children were from primarily suburban residencies. The majority of the children (87.5%) were Caucasian, and 12.5% were of Asian ethnicity (East Asian, Southeast Asian, and South Asian). Children were diverse socioeconomically though a majority was from upper/middle-class families. The majority spoke English as their primary language, but the sample included participants who spoke English as a second language. Parents received letters containing a consent form; only children with signed consent forms participated in the study. Children received a t-shirt or book for participating.

Additionally, 60 U.S.-residing adults were recruited through Amazon Mechanical Turk to obtain similarity ratings of stimuli pairs. Participants received $0.25 for completion of the task.

Design

Half of the children reasoned about people (people condition) and half about animals (animal condition). Children had to make generalizations about all people or all animals by choosing between a diverse sample (e.g., a boy and a girl) and a narrow sample (a boy and a boy). In both conditions, children were questioned about two types of properties: questions about internal biological properties (hormone) and about toy preferences (toy). There were four questions for each property (a total of eight questions), presented in block design.

Additionally, children heard either informative labels (the diverse sample had two different labels and the narrow sample had two identical labels) or no labels ("this one").

Stimuli

We created two sets of stimuli that contained either pictures of children (people condition) or pictures of animals (animal condition). The pictures of children were headshot photographs that contained children's faces and upper torsos. Twenty-four unique images of children were used (20 Caucasian children, 4 African American; 12 female, 12 male). In all images, children were smiling, face forward, and against a white background. They all appeared to be approximately 6 years old. The pictures of animals were 16 simple, semirealistic clipart-style images of familiar animals for 6-year-olds (e.g., goat, tiger, zebra, mouse). Each picture depicted the entire animal in color, against a white background. The level of detail of images was kept equivalent.

For every trial children always saw two samples: a pair of diverse people/animals (diverse) or a pair of narrow people/animals (similar). Stimuli of diverse people pairs were created using racial (e.g., a White boy and a Black boy) or gender diversity. Diverse animals pairs are animals from different taxonomic families or subfamilies. The complete set of diverse and similar pairs for people and animal conditions is presented in Table 1.

The toy questions were accompanied by full-color pictures of novel toys. The hormone questions were presented with pictures of a female scientist (for animal condition) and a female doctor (people), to provide a category-appropriate image. For each trial, the novel toy ("blick," "zav," "wug," "dax") and the novel hormone ("prolactin," "amylin," "estriol," or "cortisol") were introduced by name.

Table 1
Stimuli Sets Used in the Animal and People Domain by Block

People		Animals	
Diverse	Nondiverse	Diverse	Nondiverse
Toy block			
Boy/girl (White)	Boy/boy (White)	Lion/mouse	Lion/tiger
Boy/girl (White)	Girl/girl (White)	Monkey/horse	Monkey/gorilla
White/Black (boy)	White/White (boy)	Wolf/llama	Wolf/fox
White/Black (girl)	White/White (boy)	Goat/buffalo	Goat/sheep
Hormone block			
Boy/girl (White)	Boy/boy (White)	Zebra/tiger	Zebra/horse
Boy/girl (White)	Girl/girl (White)	Camel/gorilla	Camel/llama
White/Black (boy)	White/White (boy)	Fox/cow	Buffalo/cow
White/Black (girl)	White/White (boy)	Squirrel/sheep	Squirrel/mouse

Note. Diverse and similar pairs always appeared together. The first four sets were always used for toy questions and the second four sets were always used for hormone questions.

Procedure

Children played a game requiring them to select the "best" pair to find out if *all* people/animals "like" toy X (toy questions) or "have" hormone X. All children were presented with a block of four toy questions and a block of four hormone questions, with block order counterbalanced across participants.

For each trial, children were first introduced to a novel toy (toy image, described to children as "toys you've never heard of") or a novel hormone (doctor or scientist image, described to children as "things inside"). Next, children were told that they needed to figure out if all members of the relevant category (animals or children) "liked the toys" or "had these things inside them." They had to select between two pairs of evidence, the diverse and similar samples. Each pair of exemplars was introduced separately, and enumerated individually. The no label group heard generic labels ("this one and this one") and the label group heard labels distinguishing the diverse and narrow categories (e.g., diverse gender: "this boy and this girl"; diverse race: "this White boy and this Black Boy"). After they made their choice, they heard generic and performance-independent praise (e.g., "great!") and the next trial was initiated immediately.

Similarity Rating Study

We obtained similarity ratings for all stimuli pairs—rated by adults in the United States on Amazon Mechanical Turk. Participants rated similarity on a Likert scale ranging from 1 (*not very similar*) to 7 (*extremely similar*). For each rating, participants saw only one of the stimuli pairs, either a diverse or narrow pair. Animal and people pairs were intermixed, presented in a random order.

Results

The main dependent measure was the proportion of trials on which children selected the diverse sample (maximum $M = 1$, chance = 0.5). Children who heard informative labels did not perform differently from those who heard noninformative, generic labels, $F(1, 60) = 0.003$, $p = .957$. Hence, in all subsequent analysis we collapsed the two label conditions.

We first asked whether children's diversity-based reasoning depended on the domain (people vs. animal) and property type (within subject: hormone and toy). Indeed, children selected diverse evidence more often when reasoning about people ($M = 0.570$, $SD = 0.231$) than when reasoning about animals ($M = 0.391$, $SD = 0.249$), indicated by the main effect of domain, $F(1, 60) = 8.89$, $p = .004$, $\eta^2 = .123$. There was no overall advantage of property type, $F(1, 60) < 0.001$, $p = .971$, $\eta^2 < .001$.

We predicted that children would use diversity-based reasoning only if they believed the samples differed on their likelihood of sharing the property in question: Indeed, there was a significant interaction between domain and property type, $F(1, 60) = 6.628$, $p = .025$, $\eta^2 = .077$. Post hoc analysis confirmed our hypothesis that the advantage of the people condition only held when children were reasoning about toys, but not when reasoning about hormones. For toys, children selected diverse pairs more often in the people condition ($M = 0.633$, $SD = 0.291$, 95% CI [0.528, 0.738]) than in the animal condition ($M = 0.344$, $SD = 0.289$, 95% CI [0.239, 0.448]), $t(62) = 3.99$, $p = .0002$. But for hormones, children selected diverse pairs no more often in the people condition ($M = 0.508$, $SD = 0.321$, 95% CI [0.392, 0.623]) than in the animal condition ($M = 0.438$, $SD = 0.354$, 95% CI [0.310, 0.565]), $t(62) = 0.83$, $p = .408$ (Figure 1). Additional analysis showed that children were above chance when reasoning about people's toy preferences (binomial model: $p = .001$, $d = .291$), but below chance when reasoning about animals' toy preferences (i.e., preferring the narrow pairs; binomial $p < .0001$, $d = .297$). As predicted, children selected the diverse samples at chance level in the other two contexts: people hormones (binomial $p = .465$, $d = .016$) and animal hormone (binomial $p = .0923$, $d = .122$).

In the people condition, sample diversity could be either racial diversity or gender diversity. There was no difference between how children answered on race ($M = 0.641$, $SD = 0.121$) or gender trials ($M = 0.625$, $SD = 0.122$) for toy preferences, $t(31) = 0.205$, $p = .853$, or between race ($M = 0.469$, $SD = 0.126$) and gender ($M = 0.547$, $SD = 0.125$) for hormones questions, $t(31) = 1.15$, $p = .279$. Children were above chance for both race (binomial $p = .03$) and gender (binomial $p = .008$) in the toy preference trials.

Because 7 years is the youngest age children have been known to select diverse evidence given extra help (i.e., priming; Rhodes & Brickman, 2010), we performed a median split of the ages and compared 7-year-olds (age ≥ 6, $n = 33$, $M = 6.7$ years, $SD = 7.51$ months) and 5-year-olds (age < 6, $n = 31$, $M = 5.4$ years, $SD = 3.02$ months). Older children did not perform better overall: There was no main effect of age, $F(1, 59) = 0.121$, $p = .729$. However, there was a marginally significant interaction between domain and age group, $F(1, 59) = 3.15$, $p = .0811$. In the people condition, 7-year-olds selected the diverse samples more often ($M = 0.643$, $SD = 0.249$) than did 5-year-olds ($M = 0.514$, $SD = 0.205$), but this difference was not significant $t(30) = 1.01$, $p = .130$, $d = .565$. There was no difference in the animal condition either, $t(30) = 1.01$, $p = .320$, $d = .180$ ($M_{7\ \text{years}} = 0.442$, $SD = 0.271$, $M_{5\ \text{years}} = 0.355$, $SD = 0.214$). The full results for both age groups are presented in Table 2.

Block Order

Since children were asked about hormones and toys in blocks of four questions, to investigate a possible learning effect we repeated the mixed-design analysis with block order (first block, second

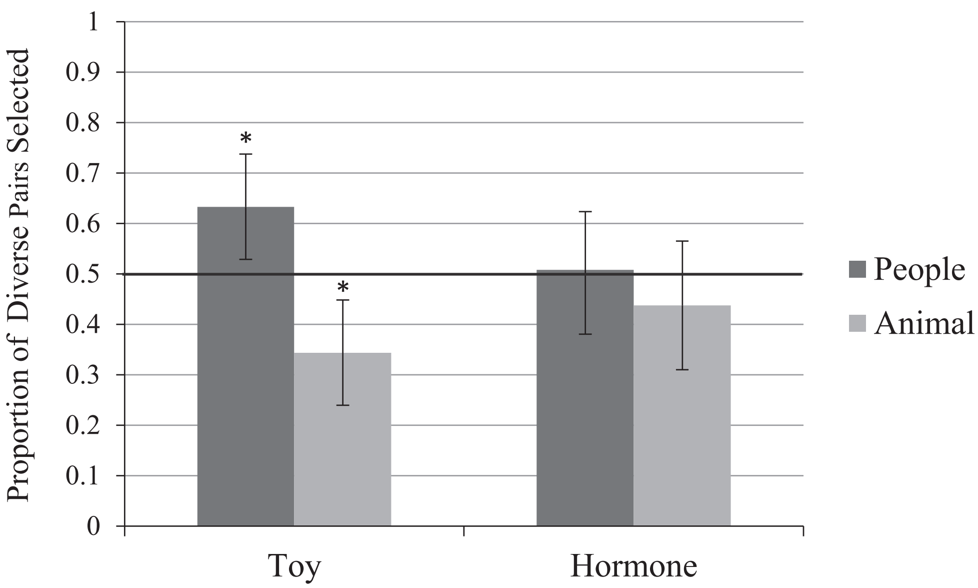

Figure 1. Proportion of diverse pairs selected. Children selected diverse evidence at above-chance level for people/toy, but below chance for animal/toy. Children were at chance for both people/hormone and animal/hormone. Error bars indicate 95% confidence intervals.
*$p < .05$ against chance.

Table 2
Comparisons to Chance by Domain, Property Type, Age, and Block

All children	*M*	*p* value	*d*	5-year-olds	*M*	*p* value	*d*	7-year-olds	*M*	*p* value	*d*
(*N* = 32) People/toy	.63	.00	.29	(*N* = 18) People/toy	.60	.06	.21	(*N* = 14) People/toy	.68	.01	.41
(*N* = 32) People/horm	.51	.47	.02	(*N* = 18) People/horm	.43	.14	.14	(*N* = 14) People/horm	.61	.07	.20
(*N* = 32) Animal/toy	.35	< .0001	.30	(*N* = 13) Animal/toy	.33	.01	.44	(*N* = 19) Animal/toy	.36	.01	.28
(*N* = 32) Animal/horm	.44	.09	.12	(*N* = 13) Animal/horm	.56	.24	.23	(*N* = 19) Animal/horm	.36	.01	.28
All children—first block				5-year-olds—first block				7-year-olds—first block			
(*N* = 16) People/toy	.69	.00	.43	(*N* = 9) People/toy	.67	.03	.38	(*N* = 7) People/toy	.71	.02	.51
(*N* = 16) People/horm	.53	.35	.06	(*N* = 9) People/horm	.42	.20	.16	(*N* = 7) People/horm	.57	.29	.15
(*N* = 16) Animal/toy	.42	.13	.15	(*N* = 7) Animal/toy	.46	.43	.07	(*N* = 9) Animal/toy	.39	.12	.21
(*N* = 16) Animal/horm	.53	.35	.06	(*N* = 6) Animal/horm	.67	.08	.38	(*N* = 10) Animal/horm	.45	.32	.10
All children—second block				5-year-olds—second block				7-year-olds—second block			
(*N* = 16) People/toy	.58	.13	.15	(*N* = 9) People/toy	.53	.43	.06	(*N* = 7) People/toy	.64	.09	.32
(*N* = 16) People/horm	.48	.45	.03	(*N* = 9) People/horm	.44	.31	.11	(*N* = 7) People/horm	.64	.09	.32
(*N* = 16) Animal/toy	.27	< .0001	.45	(*N* = 6) Animal/toy	.17	< .0001	.64	(*N* = 10) Animal/toy	.33	.02	.33
(*N* = 16) Animal/horm	.34	.01	.30	(*N* = 7) Animal/horm	.46	.43	.07	(*N* = 9) Animal/horm	.25	.00	.48

Note. Comparisons to chance using a binomial model. Children selected diverse evidence above chance only in people's toy preferences. Both 5- and 7-year-olds selected diverse evidence in the first block of people's toy preferences. *d* indicates Cohen's *d*.

block) as an additional within-subject factor. We found a significant main effect of block order, such that children selected diverse pairs significantly more often on the first block of questions ($M = 0.543$, $SD = 0.317$) than on the second block ($M = 0.418$, $SD = 0.331$), $F(1, 110)$, $p = .0174$. Consistent with our hypothesis, this first block advantage is only true for people's toy preferences ($n = 16$; $M = 0.688$, $SD = 0.266$, 95% CI [0.546, 0.829]). Interestingly, despite no overall differences in boys and girls pattern of results, girls ($n = 7$; $M = 0.857$, $SD = 0.197$) selected diverse evidence significantly more often than did boys ($n = 8$; $M = 0.531$, $SD = 0.248$) in the first block of people's toy preferences, $p = .014$, $d = .756$. Interpretations of this result, however, should be tempered by the small sample sizes of boys and girls. Children were at chance in all other contexts (Table 2).

For the second block, children were at chance in the people condition, but preferred the similar pairs in animals' toy preferences ($n = 16$; $M = 0.27$, $SD = 0.309$, 95% CI [0.101, 0.430]), and in animals' hormones ($n = 16$; $M = 0.344$, $SD = 0.328$, 95% CI [0.169, 0.518]; Figure 2). It is not clear why children favored the narrow evidence in the second block. Perhaps children thought they should implement a new strategy and found the similar animals a salient option.

Individual Strategy

To better understand the strategies of individual children in the study, children were classified as diverse selectors, similarity selectors, and chance selectors. Children were classified separately for each block, as the strategy children implemented may have changed. Diverse selectors selected diverse evidence 3–4 times out of 4, similar selectors selected diverse evidence 0–1 time out of 4, and chance selectors were those who selected diverse evidence 2 out of 4 times. The percentage of children implementing each selection strategy is outlined in Table 3.

Children's individual strategies match the parametric results: The highest proportion of diverse selectors (and lowest proportion of similar selectors) comes from people's toy preferences first block, $\chi^2(2, N = 16) = 5.09$, $p = .0784$, while the highest proportion of similar selectors (and lowest proportion of diverse selectors) comes from animals' toy preferences second block, $\chi^2(2, N = 16) = 9.52$, $p = .0086$.

Children's Verbal Rationale

At the end of the study, children were asked why they thought the group they chose was the

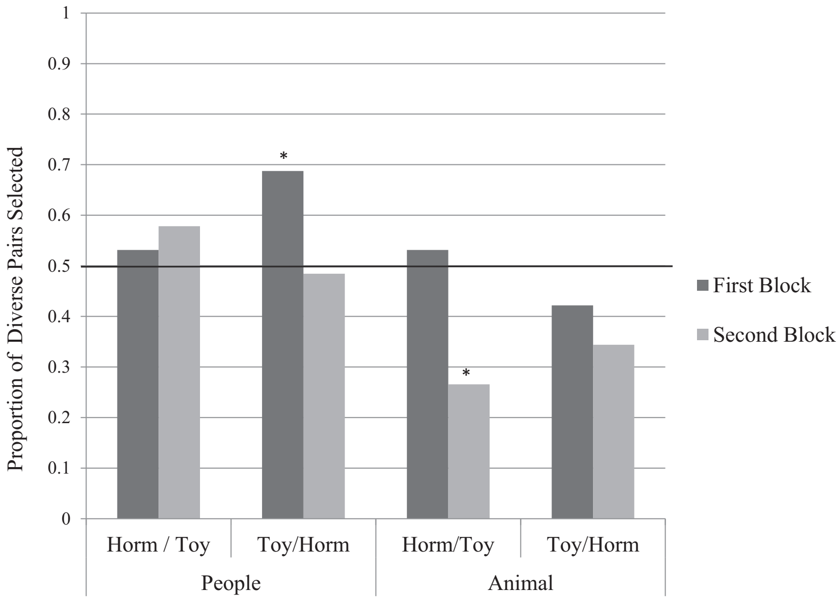

Figure 2. Proportion of diverse pairs selected by domain, property, and block. Light bar indicates mean choice of diverse samples in the first block and dark bar indicates the second block. Error bars indicate 95% confidence intervals.
*p < .05 against chance.

Table 3
Proportion of Evidence Selector Types by Domain, Block, and Property Type

	People (%)		Animals (%)	
	Block order		Block order	
Selector types	Toy [1]	Horm [1]	Toy [1]	Horm [1]
Diverse selectors	56.25	37.50	18.75	37.50
Chance selectors	31.25	37.50	50.00	31.25
Similar selectors	12.50	25.00	31.25	31.25
	Horm [2]	Toy [2]	Horm [2]	Toy [2]
Diverse selectors	31.25	37.50	18.75	12.50
Chance selectors	31.25	37.50	31.25	18.75
Similar selectors	37.50	25.00	50.00	68.75

Note. Children were classified as diverse selectors (chose 3 or 4 diverse samples), chance selectors (chose 2 diverse), or similar selectors (chose 0 or 1 diverse). Children were categorized separately for each block; each column represents one within-subject condition (n = 16).

best. Children were questioned on the fourth and eighth trial at the end of the experiment (their last toy and last hormone question). We coded these answers into five major categories: (a) *different*—children mentioned different groups as their rationale (e.g., “One is a boy and that one is a girl”; n = 15), (b) *same*—talked about sameness (e.g., “They look the same”; n = 9), (c) *liking*—chose because they liked the pair (.e.g, “I liked how they look”; n = 10), (d) *vave property*–groups they thought had the property (e.g., “They would like to play it more”; n = 15), and (e) *none*—no answer (n = 20) or said they did not know why (n = 10).

Answer *different* is most indicative of children employing diversity-based reasoning. Indeed, consistent with other analyses, children who gave the *different* justifications were more likely to be in the people condition (11/15) than in the animal condition (4/15), $\chi^2(2, N = 64) = 4.267$, $p = .0373$. Conversely, children who gave *same* rationale were more likely to be in the animal condition (7/8) than in the people condition (1/8), $\chi^2(2, N = 64) = 4.267$, $p = .0132$. No other verbal strategy differed by domain.

Similarity Ratings

We compared how adults rated diverse and similar pairs across conditions on the scale from 1 (*not very similar*) to 7 (*extremely similar*; Figure 3). As expected, adults rated diverse pairs ($M = 2.73$, $SD = 1.61$) as less similar than all similar pairs ($M = 4.50$, $SD = 1.48$), all comparisons $p < .0001$. Interestingly, there was a greater gap between diverse and similar pairs in the animal condition (mean difference = 2.20) than in the people condition (mean difference = 1.33), $t(60) = 6.29$,

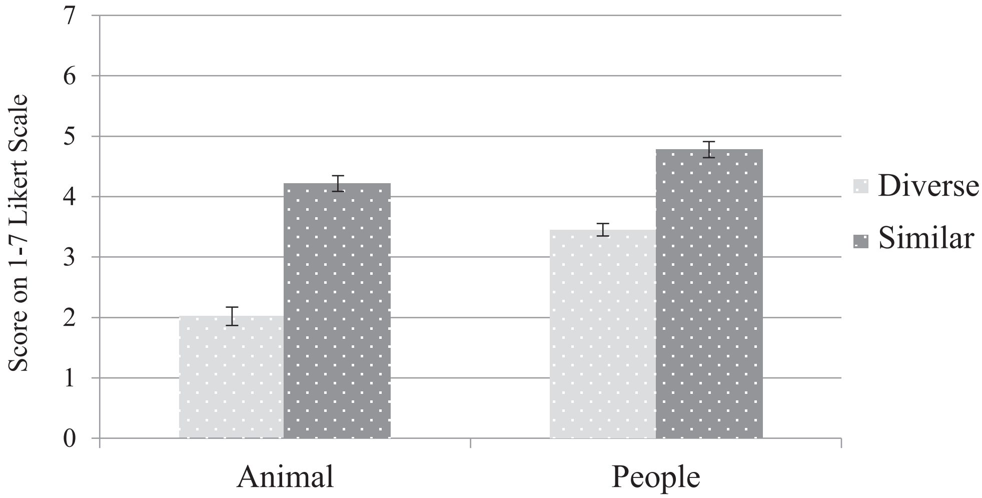

Figure 3. Similarity ratings for stimuli sets. All contrasts are significant at a $p < .0001$ level. The gap between diverse and similar pairs for animals is larger than for people, $t(60) = 6.29$, $p < .0001$, 95% CI [0.591, 1.14], $d = .521$. Error bars indicate 95% confidence interval.

$p < .0001$, 95% CI [0.591, 1.14], $d = .52$. That is, based on this perception, it is actually easier to distinguish between the diverse and similar pairs in the animal condition than in the people condition. Yet this clear distinction did not help children to use the diverse pairs in the animal condition.

Discussion

Understanding the value of diverse evidence in inductive reasoning about broader kinds is a crucial learning tool that allows young learners to build knowledge from limited initial experience. Prior studies suggest that children do not value diverse evidence until 9 years of age, suggesting an early deficit in reasoning capabilities. But here we showed that given the right context, children are quite capable of reasoning about diverse evidence. When choosing between narrow and diverse samples in an evidence selection task, even 5-year-olds reasoned that the diverse sample is more representative of people's toy preference. Consistent with prior literature, we do not see a preference for the diverse samples when children reasoned about animals' toy preference or internal properties (hormones).

Is reasoning about people special? Probably not. The more likely explanation for children's diversity reasoning success is that they believe that the members of the diverse sample are less likely to share the property in question than are members of the narrow sample. In the absence of this difference in property-sharing likelihood, applying diversity-based reasoning is moot—one might as well choose at random. Supporting this, children did not select diverse evidence when reasoning about people's internal properties—possibly because they did not assume a difference in the likelihood of people sharing hormones across race and gender. Likewise with animals, children know that different animals have different biological properties, but believe that lions/tigers' properties are roughly as different as those of lions/mice.

This belief explains why here and in prior studies with animals children did not use diversity-based reasoning. Two additional studies support this conclusion. Lo, Sides, Rozelle, and Osherson (2002) proposed that *diversity* should be replaced with a simpler model of calculating which sample is least likely to share a specific property. When children were asked about probability (which two animals would be more likely to share a property) and argument strength (which evidence is stronger for a generalization), they evaluated evidence to be stronger only if they themselves considered the animals diverse, but not when the evidence contained normatively diverse animals. Rhodes and Liebenson (2015) found that when children are explicitly provided with category variability, familiar animal stimuli or labels are sufficient to disrupt diversity-based reasoning. These and our results suggest that children understand the abstract reasoning behind diversity but evaluate animals and the distribution of their properties differently than adults do.

Children's use of diversity when reasoning about people's toy preferences cannot be explained by perceptual salience. Children restricted their selection of diverse evidence to toy preferences despite both property conditions containing equally distinguishable stimuli. Furthermore, past research using animal contrasts that are perceptually similar to racial diversity (i.e., where diversity came from black and white dogs) did not find diversity selection (Rhodes, Brickman, et al., 2008). Finally, indirect evidence from adults' ratings of stimuli suggests that the distinguishability of animals is actually *greater* than that of people. Taken together, this evidence suggests that children's diversity selection is explained conceptually (by their beliefs about category members' properties) and not perceptually.

Overall, our results suggest that young children understand the value of diverse evidence. Children disregard what adults consider diverse, not because they lack reasoning skills but precisely because they possess them. They calculate diversity on the basis of their beliefs about category members' property distributions and select diverse evidence when it is relevant. These results give us a better understanding of children's inductive reasoning in general as well as specific use of inductive reasoning in the social domain. Previous work showed that children drew inferences about individual's preferences on the basis of their category membership (Diesendruck & HaLevi, 2006). The current results extend these findings by showing that children can also use category membership to reason about how universal a preference is. Although children were not asked to make explicit generalizations here, they showed a consistent preference for varied information to chart the generalizability of a preference. We suggest that diversity-based reasoning is a valid mechanism by which even 5-year-old children can make sense of the social world—possibly employing this reasoning in trait judgments, predicting popularity or social prestige, or determining what is conventional in a given situation.

References

Diesendruck, G., & HaLevi, H. (2006). The role of language, appearance, and culture in children's social category-based induction. *Child Development*, *77*, 539–553. doi:10.1111/j.1467-8624.2006.00889.x

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. doi:10.1006/jecp.1996.2344

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, *7*, 569–592. doi:10.3758/BF03212996

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

Kinzler, K. D., Shutts, K., & Correll, J. (2010). Priorities in social categories. *European Journal of Social Psychology*, *40*, 581–592. doi:10.1002/ejsp.739

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

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgment. *Cognitive Science*, *26*, 181–206. doi:10.1016/S0364-0213(01)00066-0

López, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category-based induction. *Child Development*, *63*, 1070–1090. doi:10.1111/j.1467-8624.1992.tb01681.x

Martin, C. L., Eisenbud, L., & Rose, H. (1995). Children's gender-based reasoning about toys. *Child Development*, *66*, 1453–1471. doi:10.1111/j.1467-8624.1995.tb00945.x

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

Rhodes, M., & Brickman, D. (2010). The role of within category variability in category based induction: A developmental study. *Cognitive Science*, *34*, 1561–1573. doi: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. doi:10.1016/j.cognition.2008.03.002

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

Rhodes, M., & Liebenson, P. (2015). Continuity and change in the development of category-based induction: The test case of diversity-based reasoning. *Cognitive Psychology*, *82*, 74–95. doi:10.1016/j.cogpsych.2015.07.003

Shipley, E. F., & Shepperson, B. (2006). Test sample selection by preschool children: Honoring diversity. *Memory & Cognition*, *34*, 1444–1451. doi:10.3758/BF03195909

Shutts, K., Banaji, M. R., & Spelke, E. S. (2010). Social categories guide young children's preferences for novel objects. *Developmental Science*, *13*, 599–610. doi:10.1111/j.1467-7687.2009.00913.x