Journal of Cognition and Development

ISSN: 1524-8372 (Print) 1532-7647 (Online) Journal homepage: https://www.tandfonline.com/loi/hjcd20

Developmental Changes in the Consideration of Sample Diversity in Inductive Reasoning

Marjorie Rhodes , Susan A. Gelman & Daniel Brickman

To cite this article: Marjorie Rhodes , Susan A. Gelman & Daniel Brickman (2008) Developmental Changes in the Consideration of Sample Diversity in Inductive Reasoning, Journal of Cognition and Development, 9:1, 112-143, DOI: 10.1080/15248370701836626

To link to this article: https://doi.org/10.1080/15248370701836626

Published online: 11 Mar 2008.

Journal of Cognition and Development, 9:112–143, 2008
Copyright © Taylor & Francis Group, LLC
ISSN: 1524–8372 print / 1532–7647 online
DOI: 10.1080/15248370701836626

Developmental Changes in the Consideration of Sample Diversity in Inductive Reasoning

Marjorie Rhodes, Susan A. Gelman,
and Daniel Brickman
University of Michigan

Determining whether a sample provides a good basis for broader generalizations is a basic challenge of inductive reasoning. Adults apply a diversity-based strategy to this challenge, expecting diverse samples to be a better basis for generalization than homogeneous samples. For example, adults expect that a property shared by two diverse mammals (e.g., a lion and a mouse) is more likely to be shared by all mammals than a property that is shared by two more similar mammals (e.g., a lion and a tiger). Across four studies, we document a developmental progression in children's understanding that diverse samples provide a strong basis for generalizations, such that young children (grade 1) consistently failed to consider sample diversity within their inductive reasoning, but older children (grade 5) preferred to create diverse samples on which to base inferences about basic-level categories. These results suggest that recognizing the value of a diverse sample for inductive reasoning emerges slowly across the elementary school years.

Inductive reasoning entails making inferences from the known to the unknown, and is a crucial means of learning about and interacting with the world. Given the importance of induction for everyday thinking, a great deal of research has focused on understanding how people determine when to use information obtained from a limited sample to inform their more general beliefs (for a review, see Heit, 2000).

This research was supported by NICHD grant HD-36043 to the second author. The first and third authors were supported by funding from the Michigan Prevention Research Training Grant (NIH grant number T32 MH63057-03).

Address correspondence to Marjorie Rhodes, Department of Psychology, University of Michigan, 2024 East Hall, 530 Church Street, Ann Arbor, MI 48109-1043. E-mail: rhodesma@umich.edu

All inductive reasoning requires determining whether limited information provides a good sample on which to base more general expectations. For example, if a person learns a new fact about a bird, such as that the bird has beta cells, how far should this knowledge be generalized? Should the person then assume that all birds have beta cells? Perhaps all animals have beta cells? Alternately, should this new knowledge of beta cells be treated as applying only to this particular bird? Determining an appropriate scope of generalization is a difficult, yet critical, task. Without some means of generalizing, one would need to experience all birds individually in order to learn about them. Identifying how to generalize new knowledge adaptively may be a particularly important task for young children as they are rapidly forming and enriching their knowledge base (Heit & Hahn, 2001; Lopez, Gelman, Gutheil, & Smith, 1992).

DETERMINING WHAT MAKES A GOOD SAMPLE

A number of different models of inductive reasoning have addressed the criteria that people use to evaluate whether a given sample (referred to as the premise of an inductive argument) is strong evidence for drawing inferences about an unobserved sample (referred to as the conclusion of an inductive argument). As summarized by Heit (2000; see also Medin, Coley, Storms, & Hayes, 2003), the processes that individuals use to evaluate samples may be usefully divided into those involving *single* premises and those involving *multiple* premises. When evaluating inductive arguments involving single premises, adults generally view inductive arguments as stronger when the referent of the premise is perceived to be more similar to the referent of the conclusion (e.g., inductive arguments involving generalizations from dogs to wolves are stronger than arguments involving generalizations from dogs to horses) and when the referent of the premise is perceived to be a more typical representative of a relevant category (e.g., an argument generalizing from robins to the category of birds is viewed as stronger than an argument generalizing from penguins to birds; see, e.g., Rips, 1975; for a review, see Murphy, 2002). Developmental studies have generally revealed that children are also influenced by these factors. For example, preschool children draw more inferences about exemplars that are perceptually similar (e.g., Sloutsky & Fisher, 2004) or share category membership (Gelman & Markman, 1986), and also extend properties further when they are taught on typical examples of a category than when they are taught on atypical examples (Carey, 1985; Lopez et al., 1992).

Evaluating arguments that contain multiple premises is a more challenging task, one that involves evaluating both the characteristics of the individual premises and the characteristics of the sample as a whole. As noted by Heit (2000), evaluating the inferential power of a set of multiple premises is not the same as adding up all the features (e.g., similarity and typicality) of the individual premises. For example, consider the choice between the following two arguments:

1. Whales have gamma cells and monkeys have gamma cells. Therefore, cats have gamma cells.
2. Wolves have gamma cells and dogs have gamma cells. Therefore, cats have gamma cells.

The first argument appears stronger, despite the impression that both wolves and dogs are more typical of the inclusive category *mammals* than are whales or monkeys, and despite the impression that both wolves and dogs are more similar to cats than are either monkeys or whales.

Why does the first argument seem stronger? As described by Osherson, Smith, Wilkie, Lopez, and Shafir (1990), the first argument appears stronger than the second argument because the two premises in the first argument provide greater coverage of the category *mammals* than the two premises in the second argument. Put differently, adults expect that if two mammals as diverse as whales and monkeys both have gamma cells, then all mammals probably have gamma cells. Therefore, cats have gamma cells. In contrast, knowing that dogs and wolves have gamma cells does not seem informative about all mammals because these two examples cover only a small subcategory of mammals (e.g., canines). Therefore, despite the high level of typicality of dogs and the perceived similarity between dogs and cats, the sample of dogs and wolves does not seem as informative in this case.

The example described above reflects a preference for premise *diversity* and is consistent with general principles from the philosophy of science indicating that evidence obtained from more diverse sources more strongly supports a conclusion than evidence obtained from more homogeneous sources (for a review, see Heit, Hahn, & Feeney, 2005). A strong preference for diverse premises was found among American college students in an initial study in which participants were asked to rate the strength of arguments similar to those given above (Osherson et al., 1990). In that work, students consistently indicated that inductive arguments involving diverse premises were stronger bases of inferences than inductive arguments involving less diverse premises. This effect held both for general conclusions about the broader inclusive category (e.g., mammals) and for specific conclusions concerning particular categories (e.g., cats). Positive diversity effects among populations of American college students have been reported in a number of other studies using a variety of other methods (e.g., Gutheil & Gelman, 1997; Heit & Feeney, 2005; Kim & Keil, 2003; Lopez, 1995; Lopez et al., 1992).

PRIOR DEVELOPMENTAL EVIDENCE

There has been considerable debate about whether children also prefer diverse samples as the basis for induction. As discussed above, understanding the scope and nature of children's inductive inferences is critical to understanding how they learn about the world. Additionally, as argued by Heit (2000), examining whether children value the same sample characteristics as adults do when solving induction problems provides a window into how inductive abilities develop. He further suggests that developmental evidence should guide and constrain our evaluation of models of adult induction, such that the most successful models of induction should account not only for adults' performances, but also provide a framework for understanding the performance of children.

Several initial developmental studies, in which children were asked to make inductive inferences about animal categories based on a limited set of examples, suggested that children do not consider the degree of sample diversity when determining whether a sample provides a good basis for generalization (Carey, 1985; Gutheil & Gelman, 1997; Lo, Sides, Rozelle, & Osherson, 2002; Lopez et al., 1992). Lopez and colleagues (1992) developed picture versions of the stimuli used by Osherson et al. (1990) to assess whether 6-year-olds and 8-year-olds prefer to base inferences on more diverse sets of evidence. Children were shown pictures of a diverse set of animals that had one property and a homogeneous set of animals that had another property. Then they were asked to choose which property to extend to either another specific animal (not included in either given sample) or an animal category as a whole. In contrast to positive diversity effects found among a comparison population of college students, the authors found very limited evidence for diversity effects in children, for both specific and general conclusions. Young children did not reliably prefer to generalize the property found in the diverse set over the property found in the homogenous set.

Gutheil and Gelman (1997) hypothesized that one reason children may have had difficulty using premise diversity in an adult-like manner is that evaluating diversity requires that participants generate the relevant inclusive category (e.g., mammals) in order to assess the extent of coverage provided by the two possible samples. To address this possibility, Gutheil and Gelman (1997) developed a set of questions that asked children to make inductive inferences about a single basic-level animal category (e.g., monkeys), such that generating an inclusive category was not necessary. For example, children were shown one set of five similar looking monkeys (the homogeneous sample) that shared a particular property (orange tongues) and one set of five different looking monkeys (the diverse sample) that shared a different property (brown tongues). Children were asked to judge whether another (not pictured) monkey would most likely share the property (tongue color) with either the diverse or the homogeneous set of monkeys. Although college students reliably chose to base their inferences on the diverse set, children did not appear to consider the degree of sample diversity when making their judgments. Thus, even when the relevant inclusive category was provided (in this case, the category *monkey*), children still did not show a preference for diverse samples.

In contrast to these null findings, Heit and Hahn (2001) argue that children as young as 5 years old value sample diversity under certain circumstances. In these experiments, the authors asked participants to make inferences about everyday *objects* and their *relations to people*, as opposed to about *animals* and their *properties*. For example, children were told that three very different dolls (e.g., a china doll, a stuffed doll, and a Cabbage Patch doll) belong to Jane, whereas three very similar dolls (e.g., three Barbie dolls) belong to Molly. Participants were then asked to judge whether another target doll (e.g., a baby doll) belongs to Molly or Jane. Using this method, Heit and Hahn (2001) report strong support for diversity effects in young children, such that children were more likely to judge that the target doll belongs to the girl who owns the diverse set of dolls.

Heit and Hahn (2001) interpret these findings as evidence that young children perform diversity-based reasoning; that is, that they evaluate the extent of diversity present in a sample and prefer to base novel inferences about unknown instances on a more diverse set of evidence. Based on these findings, they conclude that children have the ability to reason about arguments involving multiple premises in an adult-like manner, and that their previous failures on diversity tasks (described above) were due to knowledge limitations (e.g., a lack of knowledge about animals or their properties; Carey, 1985), as well as limitations in their information-processing skills. Based on these findings, Heit and Hahn (2001; see also Heit, 2000) argue that models explaining the development of inductive reasoning should focus not on developmental changes in the mechanisms of inductive reasoning, but on the influence of age-related changes in knowledge.

The experiments reported by Heit and Hahn (2001) are impressive for demonstrating that young children are sensitive to sample diversity and use it as the basis of their inferences. However, several key questions remain. In the questions they presented to children, it is possible that children may have based their inferences on their perceptions of the characters, rather than on their interpretations of the sample. Specifically, these problems may be solved by appealing to what children know about people's likes and dislikes (e.g., children's knowledge that some people like all different things and some people like only one kind of thing), as opposed to their beliefs about what makes a better sample of evidence on which to base generalizations. Consistent with this hypothesis, Heit and Hahn (2001) report a failure to find diversity effects in children of the same age, using similar stimuli, when children were asked to reason about the properties of the objects as opposed to their relations to people. Thus, although these experiments demonstrate that young children can distinguish between diverse and non-diverse samples, they leave open the question of whether children believe that diverse samples provide a better basis for making generalizations than homogeneous samples.

An additional recent study with preschool children likewise reports positive diversity effects but leaves open several key questions about interpretation. Shipley and Shepperson (2006) asked children to select a sample of toys to determine which kind of toys worked properly. For example, children were shown 10 red whistles and 10 blue whistles and were asked to determine if the whistles worked properly (specifically, they were asked to decide if the whistles would make "good party favors"). Children were allowed to select two whistles to test. In these studies, children as young as 4 years typically chose to test both a red whistle and a blue whistle, as opposed to two blue whistles or two red whistles. The critical limitation in interpreting this study, however, is that children were not asked to make an inference about a larger set of whistles, including types not included in the tests (e.g., yellow whistles). Instead, they could simply have reasoned that one needs to test a blue whistle to figure out if the blue whistles work and a red whistle to figure out if the red whistles work. In this way, the task did not require children to understand that a more diverse set of whistles was better evidence for making a generalization about *novel* instances (or a larger category). Given these limitations, and the inconsistencies across previous reports, additional research is needed to examine whether and when children demonstrate an adult-like preference for diverse evidence. Furthermore, it is important to obtain insights into *why* children sometimes succeed and sometimes fail on these tasks.

THE DEVELOPMENT OF INDUCTION: CHANGES IN MECHANISMS OR CHANGES IN KNOWLEDGE?

There are two primary competing hypotheses regarding why young children fail to engage in diversity-based reasoning. As suggested by Lopez and colleagues (1992) and Gutheil and Gelman (1997), children may not have access to the adult-like mechanisms that support inductive reasoning. From this perspective, there are meaningful developmental changes in the mechanisms that support inductive reasoning, and developing adult-like strategies for evaluating the strength of inductive arguments is a developmental achievement. According to this hypothesis, young children should not demonstrate diversity-based reasoning on any type of induction problem, regardless of the domain or content of the question. Rather, young children are expected to solve induction problems by appealing to different criteria than adults do.

An alternative perspective, suggested by Heit and Hahn (2001), as well as by Shipley and Shepperson (2006; see also Hayes, Goodhew, Heit, & Gillan, 2003), is that children have access to the same mechanisms to support inductive reasoning as adults do, and that developmental differences in diversity-based reasoning relate to a number of performance-related factors. For example, children's lesser knowledge about the relevant categories (e.g., superordinate-level animal categories) may make it difficult for them to assess the degree of sample diversity that is present when reasoning about categories in a taxonomic hierarchy (e.g., Carey, 1985) or may interfere with their ability to engage in systematic reasoning (Heit & Hahn, 2001). Also consistent with this perspective, work by Medin et al. (2003) on relevance theory suggests that children may fail to demonstrate diversity-based reasoning because they view an alternate, knowledge-based strategy as more useful for solving experimental tasks.

Medin et al. (2003; see also Sperber & Wilson, 1986) propose that individuals draw inductive inferences based on the knowledge that they perceive as most relevant to the task, and not necessarily based on abstract principles (such as the diversity principle). This perspective explains the results of a series of cross-cultural studies that reported negative diversity effects among adult populations (Coley, Medin, & Atran, 1997; Coley, Medin, Proffitt, Lynch, & Atran, 1999; Lopez, Atran, Coley, Medin, & Smith, 1997). Medin and colleagues report that among populations of individuals that have a great deal of knowledge about biological categories, including the Itzaj Maya community in Guatemala, as well as experts in the United States (Proffitt, Coley, & Medin, 2000), people do not appeal to sample diversity in inductive reasoning. Rather, experts focus on their background causal knowledge related to animal behavior and/or ecological factors. For example, given questions about whether a disease shared by tapirs and squirrels (a diverse group) or by mice and rats (a homogeneous group) would be likely to spread, one participant said that the disease shared by the mice and rats would be more prevalent because these animals were likely to spread the disease more easily (Coley et al., 1999). Participants also sometimes appealed to their background knowledge about species interactions and ecological arrangements when making inferences. In these experiments, participants grounded their inferences in their prior knowledge, eliminating the need for more abstract strategies.

It is important to note that failure to attend to sample diversity in this context is not related to a failure to understand the value of diverse evidence generally. Lopez and colleagues (1997) asked adults from the same community questions such as the following, "Imagine you want to buy several bags of corn from a given person. Before buying them, this person will show you only two cobs of corn to check whether all the corn is good. Do you prefer him to show you two cobs from one and the same bag, or do you prefer him to show you one cob from one bag and another cob from another bag?" Choosing to sample one cob of corn from two different bags indicates an appreciation of the value of diverse sampling for determining whether there is support for a broader generalization (concerning the seller's corn). Using this method, Lopez and colleagues (1997) found strong support for diversity effects among Itzaj-Maya adults, suggesting that the negative diversity effects described above resulted from a preference for a knowledge-based strategy, not a lack of appreciation for diverse evidence. Relevance theory explains these findings by suggesting that adults solve inductive problems by appealing to the aspects of the task that are most relevant to the context at hand (which may vary across cultures and levels of expertise). Although relevance theory has not been specifically applied to children, the theory suggests that if children find another aspect of the task (aside from sample diversity) more salient, or if they have prior knowledge that they perceive as relevant to the task, they may fail to attend to sample diversity.

Importantly, if children do not demonstrate a preference for diverse samples either because of knowledge limitations or because they view another knowledge-based strategy as relevant to the task, then their lack of preference for diverse samples should not be interpreted as indicating that the mechanisms of inductive reasoning change developmentally. Rather, these findings would support the hypothesis that changes in knowledge underlie developmental changes in the use of sample diversity for evaluating the strength of samples. Although Heit and Hahn (2001; see also Shipley & Shepperson, 2006) conclude that the positive diversity effects reported in their experiments provide support for this hypothesis, we believe that problems with interpreting these experiments, as described above, leave open the question of how to best characterize young children's inductive reasoning.

THE PRESENT STUDIES

The goal of the present studies is to determine whether and at what age children value diverse evidence over homogeneous evidence when making inductive inferences. Toward this aim, the studies were designed to help identify which of the hypotheses described above best characterizes developmental findings related to diversity. Specifically, we aimed to design a task that focuses children as much as possible on the relative diversity present in a sample by providing children with information about samples that differ *only* in the extent to which the two samples are composed of diverse exemplars. If children demonstrate a preference for diverse samples on these simplified tasks, these findings will indicate that their difficulty in previous experiments related to a knowledge-related factor (e.g., lack of background knowledge or use of an alternate knowledge-based strategy). Alternately, if children do not prefer diverse samples on these questions, these findings will support the possibility that there are meaningful developmental changes in how young children and adults determine which samples provide a good basis for induction.

In previous studies examining diversity-based reasoning, children saw pictures of two sets of animals, learned a new property about each set, and were asked to decide which property to generalize to unobserved animals. Although presenting children with visual representations of the stimuli was intended to make the task more accessible, we suggest that this method may have distracted children from considering sample diversity. For example, children may have focused on their background knowledge about what specific kinds of animals tend to look like, and made their decision based on which animals were more familiar, typical, or attractive. Similarly, representing sets of animals in this way opens the door for a variety of other alternate strategies for answering the questions, because children may focus on other ways that the exemplars in a sample relate to one another.

We present a series of four studies examining the development of the use of sample diversity as a criterion for evaluating whether a sample is a good basis for generalization. In all studies, we manipulated diversity by varying the sampling locations. For example, children were told about one sample of four monkeys that come from a single jungle (a homogeneous sample with respect to location) and another sample of four monkeys that come from four different jungles (a diverse sample with respect to location). Children were asked to judge which sample constitutes a better basis for inferences about animals of the same basic-level category (e.g., monkeys). By relying on sampling location as our index of diversity, we eliminate a need for taxonomic knowledge. Also, in all experiments, children were presented only with simplified visual input designed to help them keep track of the story details; the actual animals were not presented visually. Thus, children were provided with no other information about the animals or the samples. Rather, they learned only that one sample contained animals from a single location, whereas the other sample contained animals from multiple locations. By varying only this single factor between the two samples, and controlling for all other information that had been presented to children in prior research (e.g., the appearance of specific animals), we aimed to increase children's attention to sample diversity.

Previous researchers have also examined diversity effects by varying sampling locations. This method is modeled loosely after research by Lopez and colleagues (1997), described above, which demonstrated that Itzaj Maya adults viewed sampling locations as an important and salient criterion for assessing the quality of samples and preferred to sample from more diverse over more homogeneous locations. As discussed by Heit and colleagues (2004), the benefit of this method of diverse sampling was first articulated by Nagel (1939), who wrote that if one wanted to determine the quality of coffee beans delivered on a ship, it would be better to inspect small samples of beans from various locations on the ship than to inspect many beans from one location. Likewise, U.S. college students view location as an important criterion for assessing sample diversity (Heit & Feeney, 2005).

In Study 1, children were asked to extend a property from either a homogeneous set of animals or a diverse set of animals to either a specific conclusion (a single animal) or a general conclusion (the category as a whole). In Studies 2 and 3, we reduced the possible influence of children's beliefs about specific properties by simply asking participants to choose whether to examine a sample drawn from one location or to examine a sample drawn from four locations. This approach permitted participants to find evidence for generalizations, without any information about where properties had previously been found. Study 4 is a control study, designed to demonstrate that even the youngest participants interpreted the stimuli as intended, such that the diverse locations were viewed as implying a more diverse sample of animals. This control study also allowed us to insure that children were able to cope with the processing demands of the task. Finally, the control study permitted us to make a direct comparison of young children's ability to detect diversity and their ability to use diversity when making inductive inferences.

STUDY 1

Study 1 was designed to follow a structure of questions similar to those asked in previous studies on diversity with children, in which children were asked to generalize a property found within either a diverse sample or a homogeneous sample (e.g., Gutheil & Gelman, 1997; Heit & Hahn, 2001; Lopez et al., 1992). We examined whether children would be more willing to generalize from a diverse sample for both specific conclusions (e.g., other category members) and general conclusions (the entire category). For all questions, the degree of sample diversity was manipulated using sampling of locations, such that children were told about one diverse sample, comprising animals from different locations, and one homogeneous sample, comprising animals from a single location. We compared reasoning across three age groups. The youngest age group included 5- to 6-year-olds, which was the youngest age group included in most previous work on diversity with children (e.g., Heit & Hahn, 2001; Lopez et al., 1992). The middle age group included 8- to 9-year-olds, also consistent with other reports on diversity (e.g., Gutheil & Gelman, 1997). We also included a comparison sample of college students.

Method

Participants

Participants were 104 students of three age groups: 32 kindergarteners and 1st graders (18 male, 14 female; *M* age = 6.5, range = 5.3–7.5); 38 3rd and 4th graders (18 male, 20 female; *M* age = 9.2, range = 8.4–10.6); and 34 college students recruited from an introductory psychology subject pool (19 male, 15 female). School-age students were recruited from an elementary school in a Midwestern university town; college students received partial course credit for participating. The sample was predominantly white.

Procedure

Each participant was tested individually in a single 5- to 10-minute session. School-age children were tested in a quiet office at their elementary school; college students were tested at an on-campus laboratory. Prior to testing, participants were told that they would be asked to look at some pictures, listen to some stories, and answer some questions. All instructions and questions were read aloud to participants and were accompanied by pictures of the described landscapes (but not of the target animals; see Figure 1).

Each participant was assigned to one of two conditions: In the general conclusion condition, participants were asked to make inferences about a category as a whole (e.g., all birds); in the specific conclusion condition, participants were asked to make inferences about a specific other category member (e.g., another bird). In both conditions, participants heard identical vignettes and saw the same pictures. For each question, participants were told about two characters who had each collected a sample of animals using different strategies: One character went to a single area and found four animals (a homogeneous sample), whereas the other character went to four different areas and found one animal in each area (a diverse sample).

For example, participants were told: "Mike went to one mountain and found four birds, and they all had tan skin under their feathers. Harry went to four different mountains and found one bird on each mountain and they all had pink skin under their feathers." The experimenter pointed to pictures of the mountains as she told the story (see Figure 1), pointing four times to the single mountain to indicate the four birds examined on the mountain, and once to each of the four mountains to indicate that one bird was examined on each mountain.

Next, participants in the specific conclusion conditions were told: "Bobby is going to *another* mountain today. Do you think he will find a bird with tan skin under its feathers like Mike found, or a bird with pink skin under its feathers like

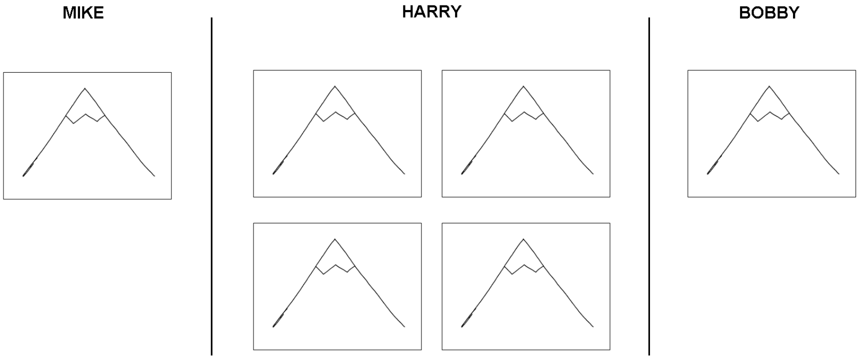

FIGURE 1 Sample pictures used in Study 1 specific conclusion condition (in the general conclusion condition, the picture on the right was omitted).

TABLE 1
Test Items Study 1

Animal	*Location*	*Property*
Monkeys	Jungle	Orange or brown tongues
Turtles	Forest	Big or small spots on stomach
Fish	Lake	Smooth or rough scales
Frogs	Pond	Croak all day or croak at the end of the day
Birds	Mountain	Have tan skin under feathers or pink skin under feathers
Butterflies	Field	Blue eyes or gray eyes

Harry found?" Participants were presented with another identical picture of a mountain, and the experimenter pointed to the relevant pictures to aid comprehension. Pre-testing with a separate set of five participants from each age group documented that children and adults correctly interpreted the target question as meaning that Bobby went to a different mountain than either of the other characters.

In the general conclusion condition, participants were told: "Of all the birds in the whole world, do you think that more of them have tan skin under their feathers like Mike found or pink skin under their feathers like Harry found?" This question was not accompanied by another picture, but pointing was used to aid comprehension.

Participants were presented with six test item sets (all questions followed the structure of the example given above; see Table 1 for a summary of the animals, landscapes, and properties used). The order of questions was randomized for each participant, the order of presentation of the two kinds of samples within each question was counterbalanced across questions, and the assignment of a particular property to a sample (e.g., tan skin in the homogeneous sample) was randomized across participants.

Control questions. After the first of the six test questions, all participants were asked a set of control questions. Participants were asked, for example, "How many mountains did Mike go to? How many birds did he find? How many mountains did Harry go to? How many birds did he find?" These questions were designed to assess whether participants understood that the different characters had examined the same number of birds and that what differed between the two samples was the locations from which these birds were drawn. If a participant failed to pass these questions, the story was retold with added emphasis that each character found a total of four animals and the questions were re-asked. Only one student (a 1st grader) repeatedly failed the control questions and was eliminated from further analyses. This indicates that children, even the youngest participants, understood the basic elements of the questions and the critical aspects of the samples.

Scoring. Across both conditions, we gave a “1” to selections of the property corresponding to the sample of one animal from each of four locations (the diverse sample), and a “0” to selections of the property corresponding to the sample of four animals from one location (the non-diverse sample). Responses were summed and divided by the total number of questions to calculate the proportion of trials in which participants favored making generalizations based on a more diverse sample.

Results

Because our outcome variable was composed of a series of dichotomous responses, we fit a series of binomial regression models to assess whether the obtained proportion of diverse responses differed from the pattern expected by chance. We found an identical pattern across both specific and general conclusion conditions. In both conditions, the proportion of diverse responses did not differ from chance for either younger children (specific conclusion: $M = .55$, $SD = .21$, $Z = .89$, $p > .3$; general conclusion: $M = .60$, $SD = .24$, $Z = 1.88$, $p > .06$) or older children (specific conclusion: $M = .51$, $SD = .33$, $Z = .19$, $p > .8$; general conclusion: $M = .52$, $SD = .36$, $Z = .18$, $p > .8$). The proportion of diverse responses for college students, however, was significantly greater than chance in both conditions (specific conclusion: $M = .84$, $SD = .21$, $Z = 6.00$, $p < .001$; general conclusion: ($M = .79$, $SD = .40$, $Z = 5.31$, $p < .001$) and, with a Bonferroni corrected level of significance, was significantly greater than the proportion of diverse responses for both younger and older children (ps < .005).

Examining participants’ response patterns further confirmed these findings. Across conditions, 75% of adult participants generalized the property found in the diverse sample on at least five of the six test questions, compared with only 21% of 3rd graders and 16% of 1st graders. The majority of 1st graders (75%) and 3rd graders (53%), but only 16% of college students, demonstrated inconsistent patterns of responding across the questions. The remaining participants extended the property found in the homogeneous sample on at least five of the six questions (9% of 1st graders; 26% of 3rd graders; 9% of college students).

Discussion

The data from both the specific-conclusion and general-conclusion conditions revealed a consistent pattern, such that neither younger nor older elementary-age children demonstrated a preference for diverse samples when making inductive inferences about either a specific conclusion (e.g., about a particular other bird) or a general conclusion (e.g., about all birds). In contrast, college students strongly favored the diverse sample for both types of generalizations (selecting the property associated with the diverse sample on 79–84% of questions). Therefore, in Study 1, we did not find evidence that children value a diverse sampling technique, consistent with some prior reports (Gutheil & Gelman, 1997; Lopez et al., 1992).

The methods used in Study 1 reduced some key confounds that may have interfered with children's reasoning in previous studies. Specifically, we eliminated the reliance on superordinate-level taxonomic categories for assessing sample diversity, by instead relying on sample locations as our index of diversity. We also provided only minimal visual input, with the aim of preventing children from basing their inferences on some unintended feature of the animals (such as similarity or typicality). These adjustments, however, did not lead children to focus on sample diversity.

We next considered that the method used in Study 1 may have led children to focus on their prior knowledge about the properties used. For example, if children previously believed that birds were more likely to have tan skin than pink skin, they may make their decision based on the property-type as opposed to an evaluation of the sample. Therefore, in the next study, we eliminated the need for participants to choose between properties.

STUDY 2

Study 2 used stimuli similar to those in the previous experiment, but asked participants to select a sample on which to base an inference (Lopez, 1995; Lopez et al., 1997), rather than to extend a particular property. In Study 2, participants were asked whether they would like to examine a sample drawn from a single area or a sample drawn from multiple, diverse areas when trying to learn about a category as a whole. Thus, this method simplified the input that participants received and reduced the likelihood that their beliefs about specific properties would be perceived as relevant. Participants had only to determine what the best sample would be for forming generalizations about a category as a whole.

Method

Participants

Participants were 184 students recruited from elementary schools and high schools in a predominantly white, working-class, Midwestern rural school district. Participants included students of four different grades: 46 1st graders (*n* = 21 male, *n* = 23 female, *n* = 2 unknown; *M* age = 7.43, range = 6.0–8.3); 48 3rd graders (*n* = 23 male, *n* = 25 female; *M* age = 9.60, range = 8.7–10.7); 52 5th graders (*n* = 30 male, *n* = 22 female; *M* age = 11.60, *SD* = .42, range = 11.0–12.7); and 38 12th grade students (*n* = 17 male, *n* = 21 female; *M* age = 18.27, range = 17.5–19.2). The two younger age groups were selected to be consistent with prior work. We also decided to include an early adolescent group (11- to 12-year-olds) in order to assess reasoning skills at the end of elementary school and to begin to determine when children demonstrate adult-like patterns of reasoning on these questions. Because the children in this study lived in a rural town, we thought that the most appropriate comparison group of adults would be 12th grade students from their local high school, as opposed to the more select group of university students sampled from in Study 1.

Procedure

Participants completed the study in a classroom setting, such that all students in the class completed the study at the same time. An experimenter read the instructions and questions aloud and provided visual support, while students followed along in individual packets and marked their answers. Students were presented with five different item sets, each asking what kind of sample they would like to collect in order to make an inference about an animal category. The animals and locations used in these test items were the same as in Study 1.

First, students were introduced to the study materials with the following introductory instructions (see Figure 2):

> This packet has pictures of different places where animals live. I'm going to ask you questions about how you want to learn about animals. Here's an example [*show poster*]. This picture shows four different forests *[point to each].* These dots *[point]* are turtles that live in the forests. So what are these? [Students answer: forests.] And what are these? [Students answer: turtles.] Great! Now pretend you're a scientist, and your job is to figure out whether turtles have big spots or little spots on their stomachs. The best way to find that out is to look at some turtles. But you can't look at *all* the turtles—you're only allowed to look at four turtles. You can choose to look at either four turtles from one forest *(point)* or one turtle from each forest *(point).* You will be asked to mark the space next to the choice that says, 'four turtles from one forest' or 'one turtle from each forest.' You can draw an X or a check mark to show your answer.

In the instructions, we used the generic form of the animal category (e.g., "… your job is to find out whether *turtles* …," as opposed to "… your job is to find out whether *these turtles* …"). This was critical because we wanted participants to understand that their goal was to learn about an animal category in general, not just the animals on the page, and prior developmental work indicates that by the age of four, children understand that generic nouns refer to general categories (e.g., Gelman & Raman, 2003; Hollander, Gelman, & Star, 2002).

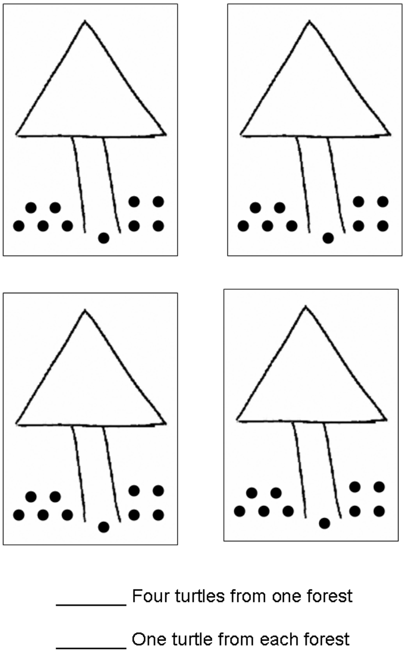
_______ Four turtles from one forest
_______ One turtle from each forest

FIGURE 2 Picture that accompanied introductory instructions, Studies 2, 3, and 4 (in Studies 2 and 4, the order of the response choices was counterbalanced, and in Study 3, the responses choices were omitted).

Next, participants received a series of five item sets, in which they were asked whether they would like to look at one animal from each of four locations or four animals from one location in order to learn about the animal category (see Table 2 for a sample question and Figure 3 for a sample picture). For each question, participants were shown one set of four landscapes and given two answer choices which were printed on the bottom of each page—for example, "If you want to look at one bird from each mountain, mark here *[point]*; if you want to look at four birds from one mountain, mark here *[point]*." The order of response choices was counterbalanced across classrooms.

Scoring

Responses in which participants chose to examine one animal from each of four locations (i.e., the diverse sample) were scored as 1 and summed. Responses in which participants chose to examine four animals from one area were scored as 0. Total diverse responses were divided by the total number of questions (5) to calculate the proportion of responses favoring a diverse sample.

TABLE 2
Sample Questions Studies 2–4

Study	*Sample Question*
Studies 2 and 4, induction condition	Here are four mountains with birds on top. You're a scientist and your job is to figure out whether birds have tan skin under their feathers or pink skin under their feathers. To find out what color skin birds have, you can look at either four birds from one mountain or one bird from each mountain. Which birds do you want to look at to help you learn about birds?
Study 3	Here are four mountains with birds on top. You're a scientist and your job is to figure out whether birds have tan skin under their feathers or pink skin under their feathers. To find out what color skin birds have, you can pick any four birds to look at. Circle the birds that you want to look at to help you learn about birds.
Study 4, appearance condition	Here are four mountains with birds on top. You're a scientist and your job is to find four birds that look really different from each other. To find four birds that look really different from each other, you can pick either four birds from one mountain or one bird from each mountain. Which birds do you want to look at to find four birds that look really different from each other?

Results

As in Study 1, because our outcome variable for each participant was composed of a series of dichotomous responses, we fit a series of binomial regression models to assess whether the obtained proportion of diverse responses differed from the pattern expected by chance. Consistent with the results of Study 1, for the youngest children (1st grade), we found that the proportion of diverse responses did not differ from the pattern expected by chance ($M = .55$, $SD = .25$, $Z = 1.58$, $p > .10$). However, the proportion of diverse responses was significantly greater than chance for 3rd graders ($M = .65$, $SD = .30$, $Z = 4.57$, $p < .001$), 5th graders ($M = .73$, $SD = .25$, $Z = 7.03$, $p < .001$), and 12th graders ($M = .71$, $SD = .34$, $Z = 5.61$, $p < .001$).

We examined age-group comparisons using a Bonferroni adjusted significance level ($p = .008$). With this criterion, we did not find that 3rd graders favored the diverse sample significantly more often than did 1st graders ($Z = 2.16$, $p = .03$); however, both 5th graders and 12th graders favored the diverse sample significantly more often than did 1st graders ($Z = 4.00$, $p < .001$, $Z = 3.31$, $p < .001$). We did not find evidence for significant differences among third, fifth, and 12th graders (ps > .06). These findings indicate that students begin to consistently prefer a diverse sample around 3rd grade.

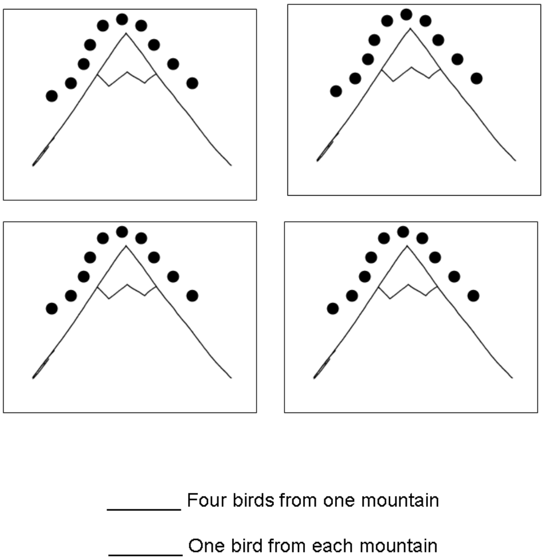
_______ Four birds from one mountain
_______ One bird from each mountain

FIGURE 3 Sample picture used in Studies 2, 3, and 4 (in Studies 2 and 4, the order of the response choices was counterbalanced, and in Study 3, the responses choices were omitted).

Examining participants' individual response patterns confirmed these findings. The percentage of participants who selected to examine the diverse sample on at least four of the five test questions increased with age (21% of 1st graders; 46% of 3rd graders; 51% of 5th graders; 60% of 12th graders), whereas the percentage of participants who responded inconsistently across questions was highest among 1st graders and lowest among 12th graders (67% of 1st graders; 41% of 3rd graders; 46% of 5th graders; 32% of 12th graders). The remainder of participants elected to examine the homogeneous sample on at least four questions (10% of 1st graders; 12% of 3rd graders; 2% of 5th graders; 16% of 12th graders).

Discussion

The findings from Study 2 demonstrate that older children (ages 8 and older) prefer a diverse sampling strategy, but that younger children do not appear to value diverse samples as the basis for making generalizations about an animal category. Study 2 revealed that children in middle childhood have an appreciation of diverse samples that was not evident in Study 1. Although the two studies used similar materials, they differed in an important way. In Study 1, participants were required to make a choice between two properties, which may have led children to expect that the specific properties should inform their decisions, thus biasing them away from considering sample diversity. In Study 2, children did not have to select between specific properties, and the performance of 3rd graders improved. The possibility that selecting between specific properties may make diversity-based tasks more difficult for children may also explain why 8- to 9-year-olds had difficulty on tasks assessing diversity in prior research (e.g., Gutheil & Gelman, 1997; Lopez et al., 1992). Thus, our findings from Study 2 suggest a more positive assessment of elementary-school-age children's abilities than suggested by prior work. We still found no evidence, however, that younger children (grade 1) value diverse evidence as a stronger basis for generalization than homogeneous evidence.

STUDY 3

Study 2 examined children's choices when they were explicitly presented with two different options for sampling, indicating that children 8 years and older recognize the value of a diverse sampling technique when the strategy is suggested to them. Because the two alternate strategies were presented to children directly, however, we cannot determine from these findings whether children in middle childhood are able to generate a diverse sampling strategy independently and, therefore, whether they are likely to prefer diverse samples in their everyday reasoning. Also, it is possible that younger children might have been negatively influenced by the presentation of two possible strategies, believing that because the experimenter suggested both strategies, both must be useful. To address these questions, in Study 3 we did not provide strategies to the children; rather, they were asked open-ended questions about which animals they would like to sample.

Method

Participants

Participants were 101 students from the same school district as the participants in Study 2; no students participated in both studies. Participants included students of the same four grades included in Study 2: 17 1st graders ($n = 7$ male, $n = 8$ female, $n = 2$ unknown; M age = 7.2, range = 6.1–8.0); 23 3rd graders ($n = 11$ male, $n = 11$ female, $n = 1$ unknown; M age = 9.4, range = 8.6–10.5); 29 5th graders ($n = 16$ male, $n = 13$ female; M age = 11.6, range = 10.5–12.8); and 30 12th graders ($n = 13$ male, $n = 17$ female; M age = 18.4, range = 17.5–19.1).

Procedure

The procedure used the same classroom-based administration techniques as were used in Study 2. Materials and instructions were also very similar, with the exception that the response choices were eliminated. Thus, after the introduction given in Study 2 (see Figure 2), children were given the following instructions:

> To find out what kind of spots turtles have, you can pick any four turtles to look at. You can pick any turtles that you want to, but you can pick only four. So, your job is to circle the four turtles that you want to look at to help you learn about turtles. You can circle any turtles on the page.

Students were then presented with five item sets; for each, they were asked to circle the dots representing the animals that they would like to look at to help them learn about an animal category (see Table 2 for a sample question and Figure 3 for a sample picture). The locations and animals in these test items were the same as in Studies 1 and 2.

Scoring

Because we used an open-ended format, there were a number of different sampling strategies open to children. However, a preliminary look at the results indicated that most responses were either diversity-based (circling one dot from each area) or homogeneity-based (circling four dots in one area), with less than 5% being other responses (e.g., two each from two areas). Therefore, each response was coded as being diversity-based, homogeneous-based, or other.

Analysis Plan

Because participants could select animals using a large number of different techniques (e.g., two animals from each area, three animals from one area and one animal from another area, etc.), the probability that a participant would select a sample of one animal from each of four locations (e.g., the diverse sampling strategy) by chance alone was quite low.[1] By contrast, in Studies 1 and 2, in which children were forced to choose between a diverse or a homogeneous sample, the probability that a diverse sample would be selected by chance alone was 50%. Therefore, Study 3 required a different analytic strategy than was used in the previous studies, and we chose to compare the proportion of diverse responses to the proportion of homogeneous responses, as opposed to comparing each to the level expected by chance.

In order to test whether the obtained proportion of diverse responses significantly differed from the proportion of homogeneous responses within grades, we conducted chi-square tests within each grade, and adjusted for the clustering of

[1] If one assumes that animals are picked by chance, then on a single question, the probability of a homogeneous response is 1 * 10/39 * 9/38 * 8/37 = .009 and the probability of a diverse response is 1* 30/39 * 20/38 * 10/37 = .11.

observations by child with the Rao-Scott Chi-square test (Rao & Scott, 1984). The results were therefore conservative in nature due to the correction of the Chi-square statistic for clustering by individual. To assess whether the proportion of diverse responses differed by grade, we conducted multinomial logistic regression analyses, with response (diverse, homogeneous, or other) as the dependent variable, homogeneous as the baseline category, and grade as an independent variable. In these analyses, we again adjusted for the clustering of responses by child. All analyses were conducted using PROC SURVEYFREQ and PROC SURVEYLOGISTIC in the SAS/STAT software package.

Results

Consistent with the findings of Studies 1 and 2, the youngest students (grade 1) did not demonstrate a preference for diverse sampling. In fact, a higher proportion of 1st graders' responses were based on a homogeneous strategy ($M = .73$) than a diverse strategy ($M = .18$; Rao-Scott $\chi2(1) = 12.14$, $p < .001$). For 3rd graders, there was no difference in the proportion of responses that were diverse ($M = .44$) or homogeneous ($M = .47$; Rao-Scott $\chi2(1) = .03$, $p > .8$). In both 5th grade and 12th grade, however, students were significantly more likely to use a diverse sampling strategy (*M*s = .70, .77) than a homogeneous sampling strategy (*M*s = .28, .08; Rao-Scott $\chi2(1) = 5.60$, $p < .02$, Rao-Scott $\chi2(1) = 30.03$, $p < .001$). Participants in both 5th and 12th grades used a diversity-based strategy significantly more than 50% of the time, *p*s < .05; in contrast, 1st graders used a diversity-based strategy significantly less than 50% of the time, $p < .05$ (see Figure 4).

As demonstrated in Figure 4, the tendency to choose a diversity-based sampling technique over a homogeneity-based technique increased with age. Specifically, 3rd grade students were 3.85 times more likely than 1st grade students (95% CI = 1.20, 12.43) to choose a diverse response than a homogeneous response. Fifth graders were 2.67 times more likely than 3rd graders (95% CI = .88, 8.10) to choose a diverse response than a homogeneous response. Twelfth graders were 3.89 times more likely than 5th graders (95% CI = 1.10, 13.79) to choose a diverse response than a homogeneous response.

Individual response patterns confirm these findings. The percentage of participants selecting diverse samples on at least four of the five test questions increased with age (6% of 1st graders; 43% of 3rd graders; 69% of 5th graders; 70% of 12th graders), whereas the percentage of participants responding inconsistently across questions was highest among 1st graders (41% of 1st graders; 17% of 3rd graders; 3% of 5th graders; 26% of 12th graders). The remainder of participants selected a homogeneous sample on at least four questions; use of this strategy declined with age (52% of 1st graders; 47% of 3rd graders; 28% of 5th graders; 3% of 12th graders).

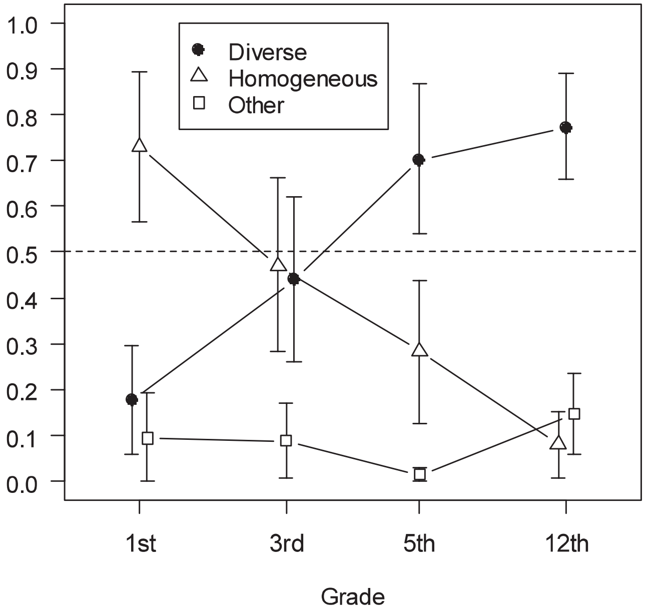

FIGURE 4 Proportion of types of responses by grade, with 95% confidence intervals, Study 3.

Discussion

The results from Study 3 indicate that children who were in 5th grade or older independently generated a diverse sampling strategy, reliably choosing to collect a diverse sample of animals to learn about an animal category as a whole. In contrast to the findings of Study 2, 3rd grade children did not reliably use this strategy, indicating that whereas 3rd grade children can recognize the value of a diverse sampling strategy when the strategy is presented to them, as demonstrated in Study 2, they cannot yet generate and apply this strategy on their own, as was required in Study 3.

Consistent with the findings from Studies 1 and 2, we found no evidence that young children (1st grade) value diverse evidence. In fact, the youngest participants were more likely to select a homogeneous sampling strategy than a diverse sampling strategy. This was the only study in which 1st graders demonstrated a significant preference for the homogeneous sample (in all other studies they demonstrated chance-level responding). Thus, it is possible that their preference for a homogeneous sample on these questions relates only to the increased performance demands of this particular task. For example, because participants had to actually circle the sample they wanted to collect, it may have been easier for them to circle four dots that were near to each other than four dots that were in different boxes.

This performance-related factor could not have influenced their performance in our other experiments, where 1st graders demonstrated chance-level responding. Therefore, we are hesitant to over-interpret the preference for the homogeneous sample documented in this study. Rather, we think the important result to note from this study is that 1st graders again failed to prefer a diverse sample, consistent with Studies 1 and 2.

STUDY 4

The aim of the final study was to eliminate a variety of alternate explanations for the consistent failure to find a preference for diverse samples among the youngest participants. Across the three previous studies, young children (1st grade) consistently failed to prefer a diverse sample as the basis for making generalizations about an animal category. However, because we used a new method to represent sample diversity in these studies than has been used in previous research, a number of alternate explanations are possible. For example, young children may have failed to assume that the animals coming from the different locations would be different from each other on any meaningful characteristics. If they did not assume that the sample coming from four locations was more diverse, then their failure to prefer this sample would not be due to a problem using diversity, but rather to a difficulty interpreting our materials. This possibility was especially important to consider given that children were not shown pictures of the animals, and the visual representations that they were shown were quite minimal. Although these aspects of the design were purposeful, in order to avoid leading children to make inferences based on visual characteristics of the stimuli, use of such stimuli was accompanied with the risk that children, especially young ones, would fail to view the sample coming from four locations as meaningfully more diverse than the sample coming from one location.

Therefore, in this final study we assessed whether 1st graders interpreted our stimuli as intended, believing that animals coming from diverse locations would have more diverse appearances than animals coming from a single location. We chose to ask about children's beliefs about the appearances of the animals because (a) prior studies have varied appearance in order to manipulate diversity (e.g., Lopez et al., 1992), and (b) all of the properties mentioned in Studies 1–3 were related to appearance (e.g., tongue or skin color). It was therefore critical to confirm that children expected the animals from diverse locations to have more diverse appearances than animals from a single location. In another condition, we sought to replicate our findings from Study 2, in which young children failed to prefer the diverse sample as a basis for generalization. Thus, Study 4 compared children's performance on a task that requires reasoning about diversity, which previous research has shown they can do effectively (e.g., Heit & Hahn, 2001; Shipley & Shepperson, 2006), with their performance on a task that requires understanding that diverse samples provide a stronger basis for generalization, which has not yet been demonstrated, and which the previous studies in this paper suggest they find difficult. In sum, in this study we test the hypothesis that 1st graders recognize that animals coming from multiple locations will be more diverse than animals coming from a single location, but that 1st graders do not prefer this diverse sample over a homogeneous sample as a basis for generalization. This study also allowed for assessment of whether children were able to follow the task and cope with the information-processing demands of our previous studies.

Method

Participants

Participants included 21 first graders (*M* age = 6.34, age range = 5.11–7.18; 8 male, 13 female; 18 white, 3 Asian-American) from the same school as the participants in Study 1. No child participated in both studies.

Procedure

Each participant was tested individually in a single 5- to 10-minute session in a quiet room at the elementary school. Prior to testing, participants were told that they would be asked to look at some pictures, listen to some stories, and answer some questions. All instructions and questions were read aloud to participants and were accompanied by pictures of the described landscapes using the same materials as were used in Studies 2 and 3.

Participants were randomly assigned to one of two conditions. In an appearance condition, children were given the following instructions (see Figure 2):

> I am going to show you some different places where animals live and ask you some questions. Here's an example *[show poster]*. This picture shows four different forests *[point to each]*. These dots *[point]* are turtles that live in the forests. (So what are these? [Student answers: forests.] And what are these? [Student answers: turtles.]) Great! Now pretend you're a scientist, and your job is to find four turtles that look really different from each other. The best way to find that out is to look at some turtles. But you can't look at *all* of the turtles—you're only allowed to pick four turtles. You can choose to pick either four turtles from one forest *(point)* or one turtle from each forest *(point)*. Your job is to find four turtles that look really different from each other. You can pick either four turtles from one forest, or one turtle from each forest.

Participants were then asked a series of five questions asking them to determine whether it would be better to look at one animal from each of four locations or four animals from one location to find animals that looked really different from each other. The animals and locations were the same as used in the previous studies (see Table 2 for a sample question and Figure 3 for a sample picture).

In the induction condition, children received instructions very similar to those given to children in Study 2. Children were told:

> I am going to show you some different places where animals live and ask you some questions. Here's an example *[show poster]*. This picture shows four different forests *[point to each]*. These dots *[point]* are turtles that live in the forests. (So what are these? Student answers: forests.] And what are these? [Student answers: turtles.]) Great! Now pretend you're a scientist, and your job is to figure out whether turtles have big spots or little spots on their stomachs. The best way to find that out is to look at some turtles. But you can't look at *all* of the turtles—you're only allowed to look at four turtles. You can choose to look at either four turtles from one forest *(point)* or one turtle from each forest *(point)*. Your job is to choose the best turtles to look at, out of all of these, to help you learn about turtles. You can look at either four turtles from one forest or one turtle from each forest.

Children were then asked a series of five questions asking them to choose to look at either four animals from one location or one animal from each of four locations to help them learn about an animal category. The animals, locations, and properties were all the same as in previous studies (see Table 2 for a sample question and Figure 3 for a sample picture).

Within both conditions, the order of answer choices was counterbalanced across participants, and the order of questions was presented in a separate random order for each participant. For both conditions, answers in which children indicated the diverse sample were scored as a 1 and answers in which children indicated the homogeneous sample were scored as a 0. Responses were summed and divided by 5 to yield the proportion of diverse responses.

Results

In the induction condition, the mean proportion of diverse responses did not differ from the number expected by chance ($M = .49$, $Z = -.14$, $p > .8$). This finding replicates our findings from previous studies indicating that 1st graders do not show a preference for choosing a diverse sample as the basis for generalizations. In contrast, in the appearance condition, the mean proportion of diverse responses was significantly greater than the number expected by chance ($M = .76$, $Z = 3.48$, $p < .0005$). Children's responses also significantly differed by condition ($Z = 2.78$, $p = .005$), with children in the appearance condition giving more diverse responses than children in the induction condition. This study demonstrates that although children expected that animals from the diverse locations would be more perceptually diverse than animals from a single location, they did not prefer to base generalizations on this more diverse sample.

GENERAL DISCUSSION

Inductive reasoning is a fundamental means of extending knowledge, and a key question is how strategies of induction compare over development. In four studies, we examined whether and at what age children believe that diverse samples provide a better basis for generalizations about animal categories than more homogeneous samples. In particular, we examined whether changes in knowledge or changes in reasoning mechanisms are responsible for age-related differences.

Evaluating whether a given sample provides a good basis for generalization is an inherent challenge in any inductive reasoning problem. Previous research suggests that adults strategically apply a diversity-based solution to this challenge, believing that a more diverse sample provides a better basis for generalization than a more homogeneous sample (Osherson et al., 1990). Thus, adults rely on the extent of the diversity present within a sample to guide their inductive reasoning and their expectations about the world. Understanding whether children also apply a diversity-based strategy informs our characterization of the strengths and limitations of children's inductive reasoning processes and the nature of change across development.

Across the present studies, we found that young children (1st grade) fail to apply a diversity-based standard for evaluating whether a sample provides a good basis for generalization. In contrast, we found that from 3rd grade onward, children recognize that a diverse sampling technique provides a better basis for inference. From 5th grade onward, we found that children are able to independently generate and apply a diverse sampling technique to investigate whether there is support for broad generalizations about basic-level categories.

The findings of Study 2 indicated a more positive description of the abilities of children in middle childhood (3rd grade) than suggested in previous reports (e.g., Gutheil & Gelman, 1997). In Study 2, when children were presented with questions asking whether they would like to look at a sample from a single location or a sample from different locations to learn about animals, third graders consistently chose to examine animals from diverse locations. These findings suggest that, under simplified conditions, 3rd graders recognize that a sample containing diverse exemplars provides a stronger basis for inference than a sample containing more homogeneous exemplars. In other words, by 3rd grade, children can successfully engage in diversity-based reasoning and appear to value similar sample characteristics as adults do when evaluating inductive strength. Their success on this task may have been facilitated by important features of the design. In particular, reducing the need to make an inference about particular properties, as was required in

Study 1 and in previous reports on diversity, may have encouraged children to attend to aspects of the sample. This understanding appears to be fragile, however, as 3rd graders did not consistently generate a diverse sample independently in Study 3. Across Studies 1–3, the results suggest an emerging ability to recognize the value of diverse evidence around 3rd grade, which becomes more robust throughout the rest of the elementary school years, as evidenced by the performance of 5th graders on our open-ended questions in Study 3. It is important to note, however, that even the oldest participants (5th graders, 12th graders, and college students) selected non-diverse samples some of the time, indicating that other factors may have influenced their reasoning on these tasks as well (see also Jacobs & Klaczynski, 2002; Klaczynski & Aneja, 2002; Klaczynski & Gordon, 1996; Medin et al., 2003).

Across all four studies, we consistently found that young children (1st grade) did not prefer to base their generalizations on more diverse samples. Study 4 indicated that these young children interpreted the test items as intended; they believed that animals coming from different locations would be more diverse than animals coming from a single location. Yet, young children did not prefer to base their generalizations on these diverse samples. This finding is consistent with a number of prior studies reporting that young children fail to value sample diversity (e.g., Lopez et al., 1992). The findings from the appearance condition of the control study, in which children correctly identified which sample was likely to be more diverse, are also consistent with other research suggesting that young children can recognize and reason about diversity (e.g., Heit & Hahn, 2001; Shipley & Shepperson, 2006). Our interpretation of prior work and the present findings suggests that although young children demonstrate skills recognizing and reasoning about diversity, and succeed on tasks that require only reasoning about sample diversity, currently available evidence does not suggest that they recognize that diverse samples provide a better basis for generalization.

A key concern in interpreting these findings is whether our youngest participants were able to reason successfully about our stimuli. Specifically, because we had purposefully presented children with stimuli that were intended only to help them track the details of the scenarios, but not to depict the diversity visually, it was possible that first graders simply did not understand that the sample coming from multiple locations would be meaningfully more diverse than a sample coming from a single location. If in fact first graders failed to expect the sample coming from diverse locations to be more diverse, or were overwhelmed by the information processing demands of reasoning about our stimuli, then their performance on our experimental tasks would not be indicative of true developmental changes in diversity-based reasoning.

The findings from Study 4, however, undermine these alternative interpretations. We demonstrated that first-grade children could reason about sample diversity using our stimuli and scenarios; they expected a sample drawn from four locations to have more diverse appearances than a sample drawn from a single location. As in our previous studies, however, they did not prefer to base inductive generalizations on this more diverse sample.

The findings from Studies 1–3, which demonstrate that young children do not demonstrate a preference for diverse samples even on simplified questions, as well as the findings from Study 4 discussed above, are counter to the proposals that young children have access to adult-like mechanisms of inductive reasoning and that age-related changes relate only to knowledge-based changes. Rather, we interpret these findings as suggesting that there may be more fundamental changes in the mechanisms that support inductive reasoning.

It is important to note, however, that this proposal does not imply that knowledge, and age-related changes in knowledge, are unimportant to induction. Indeed, prior research indicates that children have access to a range of knowledge-based strategies for performing inductive inferences, including those based on causality (Gopnik & Sobel, 2000; Kalish, 2002; Kuzmak & Gelman, 1986), perceptual similarity (Sloutsky & Fisher, 2004), category membership (Gelman, 2003), naïve theories (e.g., Keil, 1989; Wellman & Gelman, 1998), and statistical information (e.g., Jacobs & Klaczynski, 2002; Jacobs & Narloch, 2001; Piaget & Inhelder, 1975; Schlottmann, 2001). In the present experiments, however, we aimed to focus on the development of diversity-based inductive reasoning, in order to determine whether developmental changes can be attributed to knowledge-related and performance factors (Heit & Hahn, 2001), or whether changes in the underlying mechanisms may also be involved (e.g., Lopez et al., 1992; Gutheil & Gelman, 1997). We interpret the present findings as suggesting that there are important developmental changes in these mechanisms, and, therefore, that changes in both knowledge and underlying mechanisms contribute to the development of inductive reasoning.

HOW MIGHT THE MECHANISMS GUIDING INDUCTION CHANGE DEVELOPMENTALLY?

The present data do not directly address the nature of the developing mechanism; however, it is useful to consider some possibilities. Lopez et al. (1992) interpreted developmental changes related to induction, and diversity effects in particular, in terms of the similarity-coverage model described by Osherson et al. (1990). Specifically, they suggested that when children encounter a sample composed of two animals (e.g., whales and monkeys), and have to make an inference based on this sample about another animal (e.g., a cat), they have difficulty generating the inclusive superordinate-level category (e.g., mammals). Without generating the inclusive category of mammals, they cannot assess how well the two given examples cover the category of mammals, leading to negative diversity effects. The research reported by Gutheil and Gelman (1997), however, as well as the current studies, required children to make inferences about animals based on samples all at the same basic level (e.g., monkeys, birds). Therefore, these questions did not require generating an inclusive category, suggesting that this difficulty is not entirely responsible for children's failure to prefer diverse samples.

Another possibility, which would be useful to consider in future research, is that young children have difficulty distinguishing the predictive power of the individual exemplars from the predictive power of the sample as a collective whole. As described by Heit (2000) and summarized in the introduction, the strength of multiple premises is not the same as the sum of the strength of individual premises. For example, as noted earlier, a sample of whales and monkeys provides a stronger basis for inferences about cats than a sample of dogs and wolves, despite the impression that both dogs and wolves are more typical mammals and more similar to cats. However, if children evaluate each individual premise separately, as opposed to considering their group-level characteristics as a combined sample, then they would fail to attend to the aspect of the combined sample that makes the diverse sample stronger.

In the present experiments, we purposefully did not provide any information about the individual exemplars in each sample. Rather, we aimed to control for the characteristics of the individual exemplars in each sample in order to focus children as much as possible on the group-level property of diversity. In other words, participants were not provided with information about any of the individual monkeys (for example) in either sample, because we wanted them to focus only on the sampling differences (that one sample contained monkeys from a single location, whereas the other sample contained monkeys from multiple locations). In this way, although the youngest children understood that the monkeys came from different jungles, and knew that these locations were likely to contain monkeys that looked different from each other (as demonstrated in Study 4), they did not seem to have considered that together the more diverse set of monkeys had greater inferential power as a *group* than the other set of monkeys. We hypothesize that young children may, in general, evaluate samples based on the attributes of the individual exemplars in a sample as opposed to their group-level characteristics. If children fail to consider the group-level properties of a sample of exemplars, then in these experiments they would be likely to choose between the samples at random, because children were not provided with any information about the individual exemplars in either sample. This approach is consistent with previous research suggesting that young children fail to consider other group-level characteristics of multiple premise arguments, such as sample size (e.g., Gutheil & Gelman, 1997; Lopez et al., 1992). Within this framework, the present findings suggest that around third grade, children begin to recognize the value of group-level properties when evaluating the extent to which samples are informative.

FUTURE DIRECTIONS

Several questions are raised by the present studies that would be valuable to address in further research. We have suggested that the youngest children in our studies performed more poorly than young children in other studies examining diversity-based reasoning (e.g., Heit & Hahn, 2001; Shipley & Shepperson, 2006) because prior studies required children only to recognize sample diversity but not to recognize that diverse samples provide a more preferable basis for generalization. We have based this argument on the apparent differences between the present tasks and prior research, as well as on findings from our control study (Study 4), which indicated a discrepancy between young children's ability to assess sample diversity and their recognition that diverse samples provide a stronger basis for inference. It would be useful, however, to examine more directly the differences between recognizing sample diversity and selecting diverse samples as the basis of generalization, in order to better inform our characterization of inductive reasoning in early childhood.

Also, in Studies 2 and 3, we documented that children in the later years of elementary school prefer to use diverse samples as the basis for general conclusions about animal categories. We did not examine, however, whether children also use sample diversity as a standard for evaluating evidence when making inferences about specific other-category members (e.g., a particular bird). Further research examining whether children use similar sorts of criteria when making inferences about general and specific conclusions would provide a more complete picture of children's inference processes.

Finally, we have suggested that one possibility for why young children have difficulty recognizing the value of sample diversity relates to a tendency to focus on the attributes of individual exemplars in a sample as opposed to group-level properties. This possibility will be important to examine directly in future research.

ACKNOWLEDGMENTS

We are grateful to the parents, teachers, and children of the Ann Arbor Public Schools and the Delta-Schoolcraft Intermediate School District, as well as to Patrick O'Neill and Kristine Paulsen, for participating in this research.

REFERENCES

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

Coley, J. D., Medin, D. L., & Atran, S. (1997). Does rank have its privilege? Inductive inferences within folkbiological taxonomies. *Cognition, 64,* 73–112.

Coley, J. D., Medin, D. L., Proffitt, J. B., Lynch, E. B., & Atran, S. (1999). Inductive reasoning in folkbiological thought. In D. L. Medin & S. Atran (Eds.), *Folkbiology* (pp. 205–232). Cambridge, MA: The MIT Press.

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

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

Gelman, S. A., & Raman, L. (2003). Preschool children use linguistic form class and pragmatic cues to interpret generics. *Child Development, 74,* 308–325.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information bout novel causal powers in categorization and induction. *Child Development, 71,* 1205–1222.

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.

Hayes, B., Goodhew, A., Heit, E., & Gillan, J. (2003). The role of diverse instruction in conceptual change. *Journal of Experimental Child Psychology, 86,* 253–276.

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

Heit, E., & Feeney, A. (2005). Relations between premise similarity and inductive strength. *Psychonomic Bulletin & Review, 12,* 340–344.

Heit, E., & Hahn, U. (2001). Diversity-based reasoning in children. *Cognitive Psychology, 43,* 243–273.

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

Hollander, M. A., Gelman, S. A., & Star, J. (2002). Children's interpretation of generic noun phrases. *Developmental Psychology, 38,* 883–894.

Jacobs, J. E., & Klaczynski, P. A. (2002). The development of judgment and decision making during childhood and adolescence. *Current Directions in Psychological Science, 11,* 145–149.

Jacobs, J. E., & Narloch, R .H. (2001). Children's use of sample size and variability to make social inferences. *Applied Developmental Psychology, 22,* 311–331.

Kalish, C. W. (2002). Children's predictions of consistency in people's actions. *Cognition, 84,* 237–265.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: The MIT Press.

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

Klaczynski, P. A., & Aneja, A. (2002). Development of quantitative reasoning and gender biases. *Developmental Psychology, 38,* 208–221.

Klaczynski, P. A., & Gordon, D. H. (1996). Everyday statistical reasoning during adolescence and young adulthood: Motivational, general ability, and developmental influences. *Child Development, 67,* 2873–2891.

Kuzmak, S., & Gelman, R. (1986). Young children's understanding of random phenomena. *Child Development, 57,* 559–566.

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.

Lopez, A. (1995). The diversity principle in the testing of arguments. *Memory & Cognition, 23,* 374–382.

Lopez, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology, 32*, 251–295.

Lopez, A., Gelman, S. A., Gutheil, G., & Smith, E. E. (1992). The development of category-based induction. *Child Development, 63*, 1070–1090.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. L. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review, 3,* 317–332.

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

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

Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review, 97*, 185–200.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance in children*. New York: W. W. Norton & Company, Inc.

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

Rao, J. N. K., & Scott, A. J. (1984). On Chi-squared tests for multi-way tables with cell proportions estimated from survey data. *Annals of Statistics, 12*, 46–60.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior, 14,* 665–681.

Schlottmann, A. (2001). Children's probability intuitions: Understanding the expected value of complex gambles. *Child Development, 72,* 103–122.

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

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General, 133,* 166–188.

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

Wellman, H. M., & Gelman, S. A. (1998). Knowledge acquisition in foundational domains. In W. Damon (Series Ed.) and D. Kuhn & R. Siegler (Vol. Eds). *Handbook of child psychology: Vol. 2. Cognition, perception and language* (5th ed., pp. 523–573). New York: Wiley.