CHILD DEVELOPMENT

Child Development, xxxx 2016, Volume 00, Number 0, Pages 1–18

Diverse Effects, Complex Causes: Children Use Information About Machines' Functional Diversity to Infer Internal Complexity

Richard E. Ahl and Frank C. Keil
Yale University

Four studies explored the abilities of 80 adults and 180 children (4–9 years), from predominantly middle-class families in the Northeastern United States, to use information about machines' observable functional capacities to infer their internal, "hidden" mechanistic complexity. Children as young as 4 and 5 years old used machines' numbers of functions as indications of complexity and matched machines performing more functions with more complex "insides" (Study 1). However, only older children (6 and older) and adults used machines' functional diversity alone as an indication of complexity (Studies 2–4). The ability to use functional diversity as a complexity cue therefore emerges during the early school years, well before the use of diversity in most categorical induction tasks.

As adults, we constantly make inferences about unobserved aspects of objects on the basis of properties we can see. Even in the absence of detailed mechanistic knowledge, we possess skeletal framework expectations that allow us to make guesses regarding the unseen causal systems at work inside the objects we encounter (Keil, 2003). Whether we are confronted with the functions of machines or the behaviors of animals, we make inferences about inner mechanisms that can help us better understand such surface properties. Our conclusions may not always be fine-grained or even accurate, but at the broad level of inferring complexity we often have strong beliefs about the relative complexity of the hidden processes that cause observable effects. Such beliefs allow us to make decisions about how to use and classify objects, as well as whose expertise to seek if the objects need to be repaired.

Many types of surface properties give clues to the hidden complexity of human-made artifacts. Thus, an object's shape, the number and types of its external parts, and its material composition can be useful clues. Surface appearances, however, can also mislead (see Gelman & Wellman, 1991). More reliable information arises from the functional capacities of artifacts, such as speed, precision, power, efficiency, number, and diversity of functions. Functional diversity is a particularly rich and abstract clue to internal complexity: It suggests complexity in the form of multiple distinct mechanisms, or an overall complex but flexible mechanism, without suggesting specific mechanistic details or providing clear information about the kinds of mechanisms inside. Consider, for example, two machines—A and B—that look similar to each other on the outside but look different from each other on the inside. Machine A can be used to play rock songs. Machine B can be used to play rock songs and video games. Now, you are asked to guess which machine has the more complex underlying technology. You would likely choose Machine B as more complex because it has two functions rather than only one. Now, imagine you are told about two other machines. Machine C can be used to play rock songs and classical songs. Machine D can be used to play rock songs and video games.

We thank Ayotunde Ifaturoti for assistance with data collection, participant recruitment, and the drafting/preparation of stimuli; Maria Ratskevich for assistance with recruitment and the drafting/preparation of stimuli; Emily Feldstein for stimulus drawings; Elaine Bucknam for assistance with piloting; members of the Yale Cognition & Development Lab for comments on study design; and Dr. Laurie Santos, Dr. Yarrow Dunham, and three anonymous reviewers for comments on a previous draft. This research was conducted in Abington Friends School, Fayerweather Street School, Pine Point School, Renbrook School, The Country School, Wakefield Country Day School, Wooster School, Yale Peabody Museum of Natural History, and Living Laboratory® at Stepping Stones Museum for Children. We thank the generous support of the schools and museums that served as testing sites, along with the participating children, parents, teachers, staff, and administrators who made our work possible. Portions of this research were presented at the 2015 Biennial Meeting of the Society for Research in Child Development in Philadelphia, PA.

Correspondence concerning this article should be addressed to Richard E. Ahl, Department of Psychology, Yale University, 2 Hillhouse Ave, New Haven, CT 06520. Electronic mail may be sent to richard.ahl@yale.edu.

 0009-3920/2016/xxxx-xxxx
DOI: 10.1111/cdev.12613

You would likely choose Machine D as more complex: Machine C's two functions are very similar, whereas Machine D's two functions are dissimilar. Machine D has diversity in function (i.e., a wider range of functional capability), which implies more complexity in its internal mechanistic structure. You might infer that Machine D can play classical songs as well but not necessarily that Machine C can play video games.

As the previous examples illustrate, number and diversity of functions are important indications of an object's causal complexity. We can make deep inferences about the complexity of the underlying systems that enable surface properties, even if we are unable to directly see these systems ourselves. The main question we will address in this article is whether children also use functional diversity when making judgments of machines' internal complexity. Much of science works by inferring hidden properties and relations from more observable ones. Functional diversity is one source of information behind such inferences about insides not just for formally trained scientists but also for adult laypeople. It is much less clear, however, when it comes to play such a role in development. Using functional diversity to infer internal complexity requires two main abilities: the ability to appreciate the importance of insides and the ability to understand what diversity entails. Although the first appears to be early emerging, the second may take considerably more time to develop. Indeed, a substantial body of literature has shown that children have difficulty with diversity-based reasoning until the mid-elementary school years.

The Importance of "Insides"

From a young age, children believe that insides are important to agents, animals, and certain types of artifacts (Gelman & Wellman, 1991; Gottfried & Gelman, 2005; Keil, 1989; Newman, Herrmann, Wynn, & Keil, 2008; Setoh, Wu, Baillargeon, & Gelman, 2013; Simons & Keil, 1995; Sobel, Yoachim, Gopnik, Meltzoff, & Blumenthal, 2007). Eight-month-old infants expect self-propelled, agentive, and animal-like objects to have insides and are surprised if such objects are revealed to be hollow (Setoh et al., 2013). Fourteen-month-old infants more strongly relate an agentive object's behaviors to the object's internal features than its external features, expecting animated cats with stomachs of the same color, but not hats of the same color, to move in similar ways (Newman et al., 2008).

In the preschool and early elementary school years, children demonstrate explicit knowledge that the insides of natural kinds are crucial to their identities and abilities to function (Gelman & Wellman, 1991). The insides of an animal, for example, are viewed as housing its "essence" and determine the animal's functional or behavioral capabilities and categorical status, such that children expect animals from the same category to have similar insides (Gelman & Wellman, 1991). Young children know that changing the insides of animals or artifacts may also shift their category, behavior, or function, whereas changing their exteriors may not influence such properties (Gelman & Wellman, 1991; Keil, 1989; Keil, Smith, Simons, & Levin, 1998; Sobel et al., 2007). During the preschool years, children also develop specific expectations regarding the appearance of the insides that machines and animals are likely to have. For instance, by age 4, children will correctly match pictures of gears and circuit boards as belonging to machines and pictures of muscles and bones as belonging to animals (Gottfried & Gelman, 2005). However, full success on such tasks, using more subtle materials and contrasts, may only occur several years later (Simons & Keil, 1995).

In addition to making inferences about the appearance of an object's insides based on the object's categorical status, can children make inferences about the appearance of an object's insides based on the complexity of its causal effects? In other words, do children think that more complex insides will be found in machines demonstrating variable effects? One indication of such an effect was found in a study with preschoolers (Erb, Buchanan, & Sobel, 2013). In that study, children encountered two different machines. The "solid" machine's light displayed a single solid color, whereas the "variable" machine's light quickly switched from one color to the next. Children were then shown two fabricated pictures of the insides of these machines. One picture contained fewer connected parts than the other. The participants' task was to match the insides pictures to the corresponding machines. Although 3-year-olds showed no preference for one type of match, 4-year-olds showed a significant preference for matching the "complex" picture with the "variable" machine. This pattern of results was replicated when the flashing light displays were described verbally instead of shown visually. Thus, by age 4, children associate an object showing property variability with more internal complexity than an object showing property constancy.

Two possible factors may have driven 4-year-olds' inferences about the light boxes. The "variable" light might have been viewed as performing several different actions (with the flashing of each colored light as its own action) in contrast to the "solid" light, which only performed one. Alternatively, the "variable" light might have been viewed as performing a single combined action that was intrinsically more complex than that of the "solid" light. Either way, these results set the foundation for a new line of inquiry about the richness of children's inferences about unseen causes based on surface behaviors. Here, we explore children's ability to reason about more sophisticated linkages between effects and their causes by focusing on functional diversity. Can children use cues about the diversity of machines' sets of functions to determine the machines' underlying complexity when all other factors, such as number of functions and complexity of functions, are held constant? In addressing such questions, we explore the extent to which young children can make rich inferences about "hidden" causal systems, as well as provide new insights into diversity-based reasoning, a type of inductive reasoning that is generally difficult for young children.

Children's Sensitivity to Diversity

The *diversity effect* refers to the phenomenon of drawing stronger and broader inductive inferences from diverse sets of evidence than homogenous sets of evidence (Heit, Hahn, & Feeney, 2004; Kim & Keil, 2003; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990). For instance, imagine you are presented with two arguments for the claim that all animals have cesium inside (see Gutheil & Gelman, 1997; Osherson et al., 1990). The first argument is that frogs and toads have cesium inside. The second argument is that frogs and buffalos have cesium inside. You would likely deem the second argument to be stronger. The first argument only tells us that small amphibians have cesium inside, and thus this may be a property specific to their taxonomic category. The second argument tells us that a small amphibian and a large mammal have cesium inside. Because frogs and buffalos are so different, the "coverage" across frogs and buffalos is greater (i.e., more of the category of "animals" is covered by frogs and buffalos together than by frogs and toads), and so this argument provides stronger support for the claim that the presence of cesium inside is a property common to all animals. Most studies on the diversity effect in children use similar induction tasks in the domain of biology. For instance, they ask children to choose whether a diverse (e.g., different breeds of cows) or nondiverse (e.g., cows of the same breed) sample provides stronger evidence for a claim made about the entire category (e.g., all cows or all animals; Li, Cao, Li, Li, & Deák, 2009; Lo, Sides, Rozelle, & Osherson, 2002; Rhodes & Brickman, 2010; Rhodes, Brickman, & Gelman, 2008; Rhodes, Gelman, & Brickman, 2008, 2010) or choose whether a given property of a diverse sample (e.g., thin blood) or a nondiverse sample (e.g., thick blood) is more likely to be true of a novel exemplar (e.g., will a new cow of a breed not found in either set have thin blood or thick blood?) or most members of the category (e.g., most cows; Gutheil & Gelman, 1997; López, Gelman, Gutheil, & Smith, 1992; Rhodes, Brickman, & Gelman, 2008; Rhodes, Gelman, & Brickman, 2008).

Although adults readily use diversity-based reasoning (Heit et al., 2004), the evidence regarding whether young children can do so is mixed. Children younger than 8 or 9 years of age often fail to privilege diversity in inductive reasoning tasks (Gutheil & Gelman, 1997; Li et al., 2009; López et al., 1992; Rhodes, Brickman, & Gelman, 2008; Rhodes, Gelman, & Brickman, 2008); some researchers have concluded that such failures may reflect deficits in the basic cognitive skills required for mature induction (Gutheil & Gelman, 1997; Li et al., 2009; López et al., 1992). Rhodes and Brickman (2010) found that 7-year-olds were capable of favoring diverse evidence, selecting diverse animal samples as providing superior support for category-wide generalizations (e.g., claims made about all birds) than nondiverse animal samples, but only after first being primed to consider the variability within the animal category. Seven-year-olds who were not given the variability prime did not favor diverse evidence. Other studies have found that children as young as 5 can correctly use information about diversity under special conditions, such as highly pedagogical contexts (Rhodes et al., 2010), reasoning about ownership of possessions (Heit & Hahn, 2001), reasoning about people's toy preferences (Noyes & Christie, 2016), or determining the likelihood that toys function properly (Shipley & Shepperson, 2006), rather than reasoning about animals, as in most diversity-based reasoning studies. However, Rhodes, Gelman, and Brickman (2008) convincingly argue that Heit and Hahn (2001) and Shipley and Shepperson (2006) failed to test genuine diversity-based induction.

A recent study by Rhodes and Liebenson (2015) found that children as young as 5 years of age can

draw broader generalizations from diverse evidence in the domain of biology but only when reasoning about novel categories. Thus, 5- and 6-year-olds favored diverse samples as providing stronger bases for generalizations about novel categories (e.g., new animals called "modies"), but children failed to favor diverse samples for generalizations about familiar categories (e.g., birds) until the age of 9. Apparently, young children not only possess the requisite cognitive abilities for diversity-based induction, but they also possess a strong bias to view typical exemplars of a category as more informative, which interferes with diversity-based induction and makes them unlikely to favor diverse samples when reasoning about familiar categories.

Children's difficulties with diversity-based induction may be better characterized as failures to act on information about diversity rather than failures to recognize the presence of diversity. Rhodes, Gelman, and Brickman (2008) and Li et al. (2009) found that young children were capable of identifying "difference" in samples even when they did not deem diverse sets to be stronger bases for generalizations when it would have been rational to do so. Thus, the ability to recognize diversity may emerge well before the understanding of the implications of greater diversity.

Overview of Studies

Our present studies examine a different, and heretofore unexplored, type of diversity-based reasoning: that of using information about artifacts' functional diversity to make inferences about the complexity of the artifacts' insides. Here, we ask participants to reason about diversity within a single object, through its functions, rather than asking participants to reason about diversity across samples of objects. We ask participants to judge whether machines capable of executing diverse or nondiverse sets of functions have complex or simple inside parts. Thus, we are evaluating whether participants view diverse functions as stronger "evidence" for underlying mechanistic complexity than nondiverse functions. Our studies differ from most diversity-based induction tasks in a variety of ways: Our studies are about artifacts, not animals, and our key measure queries the presence of latent factors rather than category membership (see Heit & Hahn, 2001, for evidence that 5-year-olds struggle to use diversity-based reasoning regarding "hidden" properties even though they succeed in similar tasks regarding surface properties). We chose this approach because children may not have strong intuitions about taxonomically organized "machine categories," especially given the different ways that artifact and natural kind membership is construed (Keil, 1989).

Our studies investigate a novel dimension of diversity-based reasoning that still relates to the broader controversy surrounding children's competency with diversity-based induction. If children simply do not understand that diversity is informative (here, diversity supports inferences about underlying structure), they should show no preference for matching complex insides with diverse machines. However, if children are sensitive to diversity as a cue to complexity, they should preferentially match complex insides with diverse machines. Children may value information about functional diversity within a single object and use such information to infer the presence of complex, causally relevant internal parts, and they may do so at younger ages than most diversity-based reasoning studies have found. Even while lacking detailed mechanistic understanding about the workings of machines, children may still possess an abstract sense that internal complexity affords diverse functionality, whereas internal simplicity may not.

We predicted that even 4- and 5-year-old children would be sensitive to number of functions as indications of complexity and would match machines performing multiple functions with more complex insides than machines performing only one function (Study 1). However, only older children would be sensitive to diversity as an indication of complexity and match machines performing diverse functions with more complex insides (Studies 2–4). In short, we expected younger children to have difficulty integrating diversity information with judgments of internal complexity even as such a linkage is apparent to older children and adults.

Study 1

Study 1 tested whether children and adults use information about machines' number of functions, in conjunction with machines' diversity of functions, to make inferences about their internal complexity. Specifically, we tested whether participants matched machines described as performing two diverse functions with complex insides and machines described as performing one function with simple insides. This was essentially a test of sensitivity to monotonicity (i.e., sample size; see Gutheil & Gelman, 1997; Li et al., 2009; Lo et al., 2002; López et al., 1992) as a cue to complexity: Do machines

with more functions provide stronger evidence for underlying complexity than machines with fewer functions? We predicted that 4- and 5-year-olds, older children, and adults would show this pattern of responding, as the ability to use information about monotonicity when making inductive judgments is present in young children and likely precedes the ability to use information about diversity (Li et al., 2009; Lo et al., 2002). It is important to note that the findings of Erb et al. (2013) are consistent with this prediction but did not directly test whether children view number of functions as an indication of complexity. We viewed this study as necessary before addressing questions about diversity on its own. If children are insensitive to information about both diversity and monotonicity as cues to complexity, they are unlikely to be sensitive to diversity by itself as a cue to complexity.

Method

Participants

Each participant took part in only one of the studies reported here. Our final sample of children included twenty 4- and 5-year-olds (13 boys, M = 60.25 months, SD = 6.68, range = 49–69; we calculated participants' ages in months but not exact days) and twenty 6- and 7-year-olds (10 boys, M = 83.20, SD = 7.08, range = 73–94) tested in our laboratory (n = 1), children's museums (n = 15), and private schools (n = 24) in New England or the Mid-Atlantic. Thirty-three participants were White, three were Asian American, two were Black, and two were biracial. We did not collect information about socioeconomic status (SES), but given the demographic profiles of our data collection sites, we believe most children came from middle- or upper-middle-class families for all studies reported here. Data collection for all studies took place from Spring 2014 through Summer 2015. Three additional participants were excluded due to experimenter error (n = 1), comprehension difficulties and failure on the warm-up task (n = 1), or perseverative responding (n = 1; this participant always placed *complex insides* on the right side of the screen). Because we had no a priori reason to expect gender or ethnicity to affect our results, we did not run separate analyses based on these factors. In accordance with Institutional Review Board regulations, all participants had parental consent to participate and gave their personal assent.

Our final sample of adults included 20 participants (15 men). Information about age and ethnicity was not collected for adults, but all were over the age of 18 and living in the United States. Adult participants were recruited, consented, tested, and compensated online via Amazon's Mechanical Turk. Eleven had a bachelor's degree or higher. Four additional participants were excluded for failing attention and comprehension checks (n = 2), suggesting inattention or random responding, or for completing the study in a length of time that, based on both piloting and results from other participants, was extremely short (n = 2).

Materials and Procedure for Child Participants

The experimenter explained that they would be playing a game about machines: He "wanted to see what different machines looked like on the inside, so [he] opened up the front of the machines and made drawings of the parts inside the machines" (see Erb et al., 2013). He then showed two drawings of machine parts printed on small laminated cards, shown in Figure 1. We used schematized

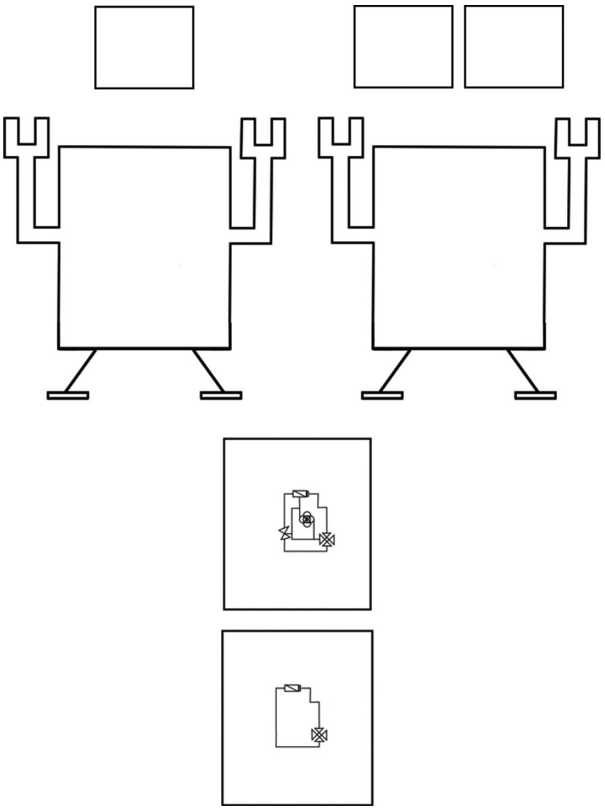

Figure 1. Schematic illustration of Study 1 stimulus presentation shown from the participant's perspective. Here, *complex insides* are shown above *simple insides* and the *one-function machine* is shown to the left of the *two-function machine*. Empty squares are displayed in place of the color photographs, representing the machines' targets, which were shown to participants.

drawings instead of realistic photos to reduce the likelihood that participants would use prior knowledge about specific machine parts to guide subsequent judgments and also to prevent the images from being perceptually overwhelming. For simplicity's sake, we will refer to the two-part picture as *simple insides* and the four-part picture as *complex insides*, although these terms were never used with participants. The experimenter showed *simple insides* first and explained that some machines "look kind of like this on the inside and have a blicket and a battery on the inside" while pointing to each part (see Erb et al., 2013, which also used contrasts of two vs. four parts). This was repeated for *complex insides*, with the addition of two other fictional parts. No further information about these specific pictures, or machines in general, was given to the participants.

After introducing the insides pictures, the experimenter placed an Apple iPad in front of the participants and placed the insides pictures, oriented vertically, between the iPad and the participants. Whether *complex insides* was closer to the participant or the iPad was randomly determined for each participant. The Qualtrics Offline iPad app was used for stimulus presentation and data collection. As extra checks, we also live-coded participants' responses and made video recordings of participants whose parents gave their permission.

The study began with a warm-up task designed to draw participants' attention to the perceptual differences between *simple insides* and *complex insides* and acclimate participants to placing picture cards on the iPad screen. Participants were shown two schematized outlines of robot-like machines on the iPad, oriented vertically. One machine's insides resembled *simple insides*, and the other machine's insides resembled *complex insides*. Participants were instructed to place *simple insides* and *complex insides* on top of the appropriate machines. This was the only task in our study for which instructive feedback was given. In all studies reported here, all but five participants made the correct match on their first attempt; three of these participants made the correct match on a repeat attempt, demonstrated an understanding of the initial mistake, and were thus included in the sample.

Each Conceptual Matching task item consisted of two outlines of robot-like machines along with pictures of different objects shown above the outlines. (The machine outlines were identical for both machines in a given pair. Each pair had unique outlines different from those of the other pairs.) The experimenter pointed to the machines and explained, "they have similar parts on the outside but do different things and look different on the inside." The participant's job was to indicate what each machine looks like on the inside by placing *simple insides* and *complex insides* on the appropriate machines. With the exception of clarifying instructions on *how* to do the Conceptual Matching task (e.g., placing the pictures on the iPad), no specific feedback was given. All participants correctly understood that each picture could only be placed on one machine in each pair. The Conceptual Matching task consisted of six items. Examples of a single item from all four studies are shown in Appendix A, and a complete list of items is included in Appendix S1. One machine in each pair, which we will refer to as the *one-function machine*, performed one function ("this machine can make cupcakes"), while the other machine, the *two-function machine*, performed two functions (the *one-function machine's* function plus a new one, e.g., "this machine can make cupcakes and soups"). The functions were stated verbally by the experimenter as well as conveyed visually, with the targets of the machines' functions displayed above the machines. Using Qualtrics, the order in which the six items were administered was randomized for each participant. Whether the *two-function machine* appeared on the left or right of the screen was randomized for each participant and each item. Whichever machine appeared on the left was discussed first. Similar randomization procedures were used for all developmental studies reported here.

Following the Conceptual Matching task, participants completed a final one-trial check on attention and matching abilities, modeled after Erb et al. (2013), in which participants matched cards depicting colored shapes with images displayed on the iPad. All participants in Studies 1–4 succeeded. At the end of the study, participants were thanked, debriefed (among other things, we explained that the machines were not real), and given a small prize. The study flow for tasks in all studies is shown in Appendix B.

Materials and Procedure for Adult Participants

The overall procedure was similar with adult participants, with some modifications to make the task age appropriate and amenable to self-administration without a live experimenter. Most notably, we added new attention check questions suitable for online studies, modified the instructions to make the task more plausible for adults (we added the explanation that the pictures were "extremely

simplified drawings" and did not name the individual parts), and displayed instructions on-screen via text. The order in which the Conceptual Matching task items appeared was randomized, although the *two-function machine* appeared first and on the left for three of the items and second for the other three items for all participants (this feature was fully randomized for the child participants). So as not to encourage the adult participants to "think like kids," we did not tell them that the study was also being conducted with children until the debriefing section at the end of the study.

Results

For each of the six items on the Conceptual Matching task, participants were given a 0 if they matched *simple insides* with the *two-function machine* and a 1 if they matched *complex insides* with the *two-function machine*. Total scores could range from 0 to 6, with a score of 0 indicating a belief that internal simplicity is associated with greater functional capacity, a score of 3 indicating at-chance, random, or idiosyncratic response strategies, and a score of 6 indicating a belief that internal complexity is associated with greater functional capacity, which we viewed as a "correct" pattern of responding. We conducted separate one-sample t tests for each age group, comparing mean scores with the at-chance score of 3.0, to assess whether participants showed a preference for matching *complex insides* with the *two-function machines*. Four- and 5-year-olds, $t(19) = 4.36$, $p < .001$, Cohen's $d = 0.975$, 6- and 7-year-olds, $t(19) = 12.86$, $p < .001$, $d = 2.875$, and adults, $t(19) = 13.81$, $p < .001$, $d = 3.089$, all matched *complex insides* with the *two-function machines* significantly more often than would be expected by chance. Item-specific success rates for all studies are reported in Appendix S2.

An analysis of variance (ANOVA), with Conceptual Matching task score as the dependent measure and age as the between-subjects factor, revealed a significant main effect for age, $F(2, 57) = 5.75$, $p = .005$, $\eta_p^2 = .168$, whereas Levene's test indicated heterogeneity of variance ($p < .001$). A follow-up comparison showed that 4- and 5-year-olds' scores ($M = 4.50$, $SD = 1.54$) were significantly lower than those of 6- and 7-year-olds ($M = 5.55$, $SD = 0.89$), $t(38) = 2.64$, $p = .013$, $d = 0.858$, with significance-level adjustments to account for variance heterogeneity. Six- and 7-year-olds and adults ($M = 5.55$, $SD = 0.83$) had identical mean scores.

Participants were divided into three groups based on whether their scores revealed a strong preference for matching *simple insides* to the *two-function machines* (score of 0 or 1: Group A), a weak preference in either direction or at-chance responding (score of 2, 3, or 4: Group B), or a strong preference for matching *complex insides* to the *two-function machines* (score of 5 or 6: Group C). Results are displayed in Table 1 for this and all other studies.

Discussion

Consistent with our predictions, participants of all age groups showed a significant preference for matching *complex insides* with the *two-function machines*. The score distribution shown in Table 1 indicates that the improved performance of the 6- and 7-year-olds relative to younger children can be attributed to 6- and 7-year-olds' decrease in Group B responses (i.e., random or idiosyncratic responses) as opposed to a decrease in Group A responses (i.e., responses consistent with a belief that internal simplicity is associated with greater functionality), as no participants exhibited a Group A response pattern.

Children attended to the causal relevance of insides and associated internal complexity with the

Table 1
Participants' Conceptual Matching Task Response Patterns for Each Study and Age Group

	Group A (% scoring 0 or 1)	Group B (% scoring 2, 3, or 4)	Group C (% scoring 5 or 6)
Study 1			
4- and 5-year-olds	0	45	55
6- and 7-year-olds	0	15	85
Adults	0	10	90
Study 2			
4- and 5-year-olds	10	70	20
6- and 7-year-olds	5	55	40
8- and 9-year-olds	0	25	75
Adults	0	10	90
Study 3			
6- and 7-year-olds	5	45	50
8- and 9-year-olds	5	35	60
Adults	0	5	95
Study 4			
4- and 5-year-olds	0	80	20
6- and 7-year-olds	5	50	45
Adults	0	20	80

Note. This table shows the percentage of participants who showed a strong preference for matching *complex insides* to the *one-function* or nondiverse machine (Group A: score of 0 or 1), no strong preference for either type of match (Group B: score of 2, 3, or 4), or a strong preference for matching *complex insides* to the *two-function* or *diverse machine* (Group C: score of 5 or 6), for each study and age group.

ability to do a greater number of things (i.e., greater functionality). Children ages 4 and older seemed to be sensitive to information about monotonicity, at least when combined with information about diversity: The ability to perform two functions is stronger evidence for underlying complexity than the ability to perform only one function. However, Study 1 does not provide evidence that children view complexity as correlated with functional diversity per se. When making their complexity matching judgments, it is unclear whether participants were attending to the diversity of the *two-function machine*'s set of functions above and beyond the number of functions in its set. Moreover, participants' apparent successes at this task could be attributed to a simple perceptually based strategy of matching "more" to "more" (relative to *simple insides*, *complex insides* had more parts and the *two-function machine* was associated with pictures of two objects and took longer to explain), as opposed to a strategy that incorporates the understanding that insides are causally relevant to functionality. Although we believe that mere perceptual matching is unlikely to fully explain our results (many children spontaneously explained that machines with "more complicated parts" are "better," "smarter," and can "do more things"), surface features of our task may have inflated participants' performance. We modified our task in Study 2 to directly test whether children use information about functional diversity to make internal complexity judgments.

Study 2

In Study 2, the Conceptual Matching task items conveyed contrasts of diversity; both machines in each pair performed an equal number of functions, but only one machine performed diverse functions. We also added a new task to test whether participants were capable of identifying the correct machine as having diverse functionality. Because previous studies have found limited evidence that children below the age of 8 privilege information about diversity in inductive reasoning tasks, we expanded our age range a priori to include 8- and 9-year-olds.

Method

Participants

Our final sample of children included twenty 4- and 5-year-olds (13 boys, M = 60.70 months, SD = 6.67, range = 49–71), twenty 6- and 7-year-olds (12 boys, M = 83.65 SD = 7.04, range = 73–95), and twenty 8- and 9-year-olds (12 boys, M = 108.55, SD = 6.77, range = 96–117), tested at children's museums (n = 25) and private schools (n = 35). Fifty participants were White, three were Asian American, two were Black, two were biracial, one was Latino, and one was American Indian. Ethnicity information was unavailable for one participant. Two additional participants were excluded due to severe inattention (n = 1) or perseverative responding (n = 1). Our final sample of adults included 20 participants (10 men). Twelve had a bachelor's degree or higher. Four additional participants were excluded for failing attention and comprehension checks (n = 2) or for extremely short completion times (n = 2).

Materials and Procedure

The materials and procedure were similar to those in Study 1 with the two notable exceptions of modifications to the machines' set of functions and the addition of the Different Introduction task and the Different task at the end of the study. We included the Different task to determine whether potential failures on the Conceptual Matching task could be explained by a failure to identify diversity, particularly in our youngest participants.

In the Conceptual Matching task, the paired contrasts were now between machines that each performed two functions. However, in each pair, only one machine had diversity in functionality. We will now refer to the two machine types as the *diverse machine* and the *nondiverse machine*. The *diverse machine*'s functions were identical to those in Study 1. The nondiverse machine performed functions on two objects, although the two objects were of the same kind (e.g., two cupcakes). The targets of the *diverse machine*'s functions suggested diversity (the process of making cupcakes is quite different from making soups, and thus this machine's functional "coverage" is extensive), whereas the targets of the nondiverse machine suggested a lack of diversity. To reduce the likelihood that participants could simply use visual information about the perceptual diversity or complexity of the *diverse machine*'s target objects to inform their internal complexity judgments, without considering contrasts in the kinds of objects, we chose new images that differed in camera angle, color, and so on., for the second target associated with the nondiverse machine, compared to the first target, which was common to both machines.

Prior to starting data collection, we pretested the functions used in our items by having a separate group of adults (n = 72) rate how technologically advanced a machine would need to be to perform each individual function (e.g., making cupcakes). Pretesting details are explained in Appendix S3. In our final list of items, the functions in a given item received similar ratings from adult participants (all ps > .10). For instance, "making cupcakes" was rated as requiring a similar level of technological sophistication as "making soups." Thus, participants' performance in Study 2 would be unlikely to be driven by an impression that an individual function was more complex than the others in the given item, and thus matching *complex insides* to the *diverse machine* that made cupcakes and also soups can be best attributed to the diversity of that machine's set. Although it is possible that children's impressions of each function's complexity may diverge from those of adults, we believe they are unlikely to do so in ways that would systematically affect our overall pattern of results.

The Different Introduction task, included in Appendix S4, was designed to familiarize children with making decisions about similarities and differences and create a transition from the Conceptual Matching task to the Different task. In one of the questions, the experimenter told participants about a child who sometimes goes to school in a van and sometimes in a car, and another child who sometimes goes on a bike and sometimes in a train, and asked participants to identify which child gets to school in ways that are different from each other. The other question had a similar structure but was about food choices. The left–right position of the "different" child, as well as which of the questions was asked first, was randomized across participants. One-sample t tests comparing each age groups' total scores, which could range from 0 (*both incorrect*) to 2 (*both correct*), with the at-chance score of 1, found that 4- and 5-year-olds (p = .004), 6- and 7-year-olds (p < .001), and 8- and 9-year-olds (p < .001) all exceeded at-chance performance. (Participants in Studies 3 and 4 performed similarly well, and because these results are not of primary interest, we will not discuss them further.)

The Different task tested whether participants were capable of identifying the *diverse machine* as the machine with more functional diversity. Participants were reintroduced to the machines shown previously and told that "some do things that are similar, and some do things that are different." The six Conceptual Matching task items were displayed again in a new randomized order. For each item, the experimenter restated both machines' functions and asked participants to point to the machine that does "different" things. Importantly, the structure of the Different Introduction task items diverged from that of the Different Task items: The Different Introduction task involved reasoning about people and involved a different kind of contrast than the Different task (one target was common to both machines in the Different task but not to the children in the Different Introduction task). Thus, the Different Introduction task did not train participants on the exact type of response pattern required for success in the Different task.

Results

For each *conceptual matching* task item, participants were given a 0 if they matched *simple insides* with the *diverse machine* and a 1 if they matched *complex insides* with the *diverse machine*. As shown in Figure 2, total scores could range from 0 to 6, with a score of 6 consistent with a belief that internal complexity is associated with functional diversity, which we viewed as a "correct" pattern of responding. We conducted separate one-sample t tests for each age group, comparing mean scores with the at-chance score of 3.0, and found that 4- and 5-year-olds' scores were no different from chance, $t(19) = 1.58$, $p = .13$, $d = 0.353$. Six- and 7-year-olds, $t(19) = 3.61$, $p = .002$, $d = 0.807$, 8- and 9-year-olds, $t(19) = 8.46$, $p < .001$, $d = 1.891$, and adults, $t(19) = 15.77$, $p < .001$, $d = 3.527$, all matched *complex insides* with the *diverse machines* significantly more often than would be expected by chance. An ANOVA with Conceptual Matching task score as the dependent measure and age as the between-subjects factor revealed a significant main effect for age, $F(3, 76) = 12.08$, $p < .001$, $\eta_p^2 = .323$, although Levene's test revealed heterogeneity of variance (p = .04). Follow-up Bonferroni-adjusted comparisons (resulting in an adjusted α of $p < .017$) revealed that 6- and 7-year-olds' scores were not significantly greater than those of 4- and 5-year-olds, $t(38) = 1.64$, $p = .11$, $d = 0.531$, whereas 8- and 9-year-olds' scores were significantly greater than those of 4- and 5-year-olds, $t(38) = 4.45$, $p < .001$, $d = 1.443$, and marginally greater than those of 6- and 7-year-olds, $t(38) = 2.45$, $p = .02$, $d = 0.796$, with significance-level adjustments to account for variance heterogeneity.

For each Different Task item, participants were given a 0 if they chose the nondiverse machine as the machine that "does different things" and a 1 if they chose the *diverse machine*. One 4-year-old

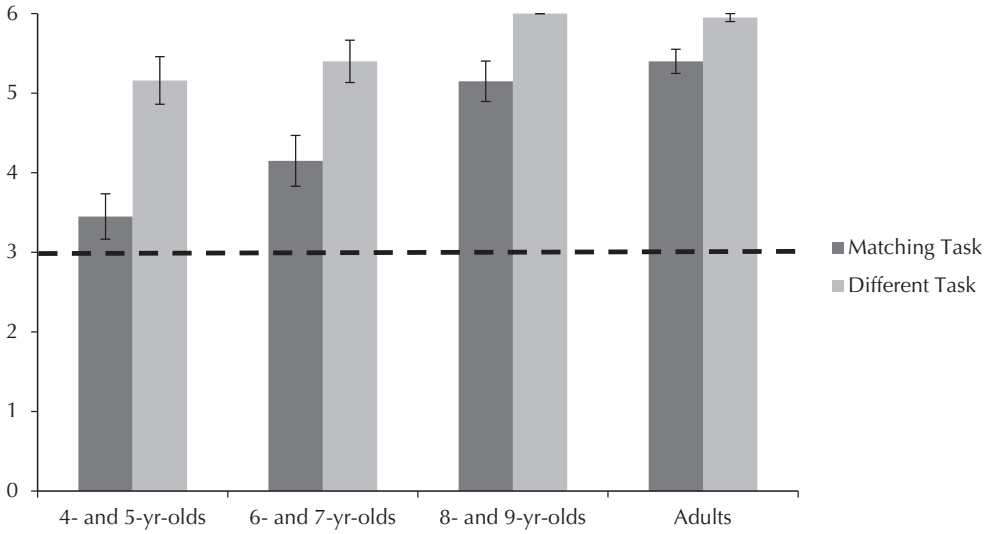

Figure 2. Mean scores in the Conceptual Matching task and Different task in Study 2. For each task, total scores could range from 0 to 6, with a score of 3 (shown with a dotted line) indicating at-chance performance. Four- and 5-year-olds' Matching task scores were not significantly different from the at-chance score of 3.0. Their Different task scores, as well as all other age groups' scores on both tasks, significantly exceeded chance. Error bars indicate 1 *SEM* in either direction.

discontinued the Different task due to boredom, reducing the sample to 19 participants for the 4- and 5-year-old age group for this section alone. We conducted separate one-sample t tests for each age group and found that 4- and 5-year-olds, $t(18) = 7.22$, $p < .001$, $d = 1.657$, 6- and 7-year-olds, $t(19) = 9.04$, $p < .001$, $d = 2.021$, 8- and 9-year-olds (no participants scored < 6.0), and adults, $t(19) = 59.00$, $p < .001$, $d = 13.193$, all correctly identified the *diverse machine* significantly more often than would be expected by chance.

Discussion

Results from Study 2 indicate that, by the ages of 6 and 7, children can associate functional diversity with underlying complexity at rates exceeding chance. Conceptual Matching performance improved with age, this time approaching adult-like levels among 8- and 9-year-olds but not 6- and 7-year-olds. Four- and 5-year-olds' mean Conceptual Matching score was no better than chance, suggesting that they had difficulty using information about diversity in function to make inferences about underlying mechanistic complexity. However, their mean Different task score was well above chance. In fact, because the Different task was the final section of a study that often took more than 10 min to complete, and children's attention may have waned over the course of the session, we may have underestimated their true abilities.

Results from Study 1, as well as those from Erb et al. (2013), indicate that 4- and 5-year-olds associate internal complexity with greater efficacy and functionality. Different task results show that the children of this age can identify the *diverse machine* as such, suggesting that their Conceptual Matching failure cannot be attributed to a total inability to see diversity when it is present. Indeed, the mean Conceptual Matching task score of 4- and 5-year-olds who scored a 5 or 6 on the Different task ($M = 3.36$, $SD = 1.45$) was similar to the mean conceptual matching task score of all 4- and 5-year-olds ($M = 3.45$, $SD = 1.28$). Thus, although 4- and- 5-year-olds showed competency at some components that are likely necessary for high Conceptual Matching task scores, they did not integrate these components in a manner that allowed them to perform well. Our findings of a disconnect between Conceptual Matching task and Different task performance echo the results of Rhodes, Gelman, and Brickman (2008), who found that first graders did not select diverse over nondiverse samples as providing a stronger basis for generalizations, although children of this age were able to correctly identify the diverse samples as "more different" (see also Li et al., 2009; Shipley & Shepperson, 2006).

In Studies 3 and 4, we explore first the extent of older children's competencies and then younger children's weaknesses. One alternative interpretation of the older children's Conceptual Matching success in Study 2 focuses on the potential obviousness of the

nondiverse contrast cases. According to this view, older participants did not view the nondiverse machine as performing two separate, albeit nondiverse, functions (e.g., making cupcakes with dark frosting and making cupcakes with light frosting), but rather as performing a single function (e.g., making cupcakes), and therefore Study 2 tested monotonicity rather than diversity. Thus, in Study 3, we investigate children's performance with a more difficult test of sensitivity to diversity.

With respect to younger children's performance, 4- and 5-year-olds apparently did not spontaneously attend to diversity in the Conceptual Matching task or simply did not view the type of diversity we conveyed as relevant to judgments of internal complexity. We chose to convey diversity by varying functional targets, rather than the verbs themselves, so we could convey the diversity contrasts visually (it is difficult to convey the meaning of verbs in a visual manner), thus minimizing memory demands and facilitating comparisons. However, at least in the context of the Conceptual Matching task, younger children may have struggled to understand that different targets suggested qualitatively different types of functions because the same verbs were used for both machines. In Study 4, we explore the possibility that young children may view diversity as an indication of underlying complexity if the diversity is conveyed through verbs as well as noun targets.

Study 3

Perhaps older children were only using a shallow "different name" heuristic to infer diversity. To assess that possibility, in Study 3, instead of having the nondiverse machine perform its two functions on the same lexical item, it performed functions on two distinct lexical items. As 4- and 5-year-olds' Conceptual Matching task performance in Study 2 was no better than chance, we did not test children of this age in Study 3.

Method

Participants

Our final sample of children included twenty 6- and 7-year-olds (11 boys, $M = 83.25$, $SD = 6.06$, range = 73–95) and twenty 8- and 9-year-olds (9 boys, $M = 108.05$, $SD = 7.98$, range = 97–119) tested at children's museums ($n = 14$) and private schools ($n = 26$). Thirty-one participants were White, two were Asian American, one was Black, and one was biracial. Ethnicity information was unavailable for five participants. Three additional participants were excluded due to experimenter error ($n = 2$) or equipment problems ($n = 1$). Our final sample of adults included 20 participants (13 men). Twelve had a bachelor's degree or higher. Six additional participants were excluded for failing attention and comprehension checks ($n = 3$) or extremely short completion times ($n = 3$).

Materials and Procedure

The materials and procedure were identical to those in Study 2 except that, in Study 3, the nondiverse machine's targets were two distinct lexical items. For instance, the nondiverse machine made cupcakes and muffins rather than two cupcakes. For three items (items 1, 5, and 6), the functional target unique to the nondiverse machine was a synonym of the target common to both machines. For the other three items (items 2–4), the functional target of the nondiverse machine was closely related to the common target but was not a synonym. The functional targets unique to Study 3 were pretested at the same time as the Study 2 items with the same participants, as explained in Appendix S3. The functions of the nondiverse machine and the functions unique to the *diverse machine* received similar technological sophistication ratings from adult participants (all ps $> .10$).

Results

Separate one-sample t tests for each age group, comparing mean Conceptual Matching task scores with the at-chance score of 3.0, found that 6- and 7-year-olds, $t(19) = 3.56$, $p = .002$, $d = 0.796$, 8- and 9-year-olds, $t(19) = 5.71$, $p < .001$, $d = 1.276$, and adults, $t(19) = 12.70$, $p < .001$, $d = 2.839$, all matched complex insides with the *diverse machines* significantly more often than would be expected by chance, as shown in Figure 3. An ANOVA with Conceptual Matching task score as the dependent measure and age as the between-subjects factor revealed a significant main effect for age, $F(2, 57) = 6.40$ $p = .003$, $\eta_p^2 = .183$, although Levene's test revealed heterogeneity of variance ($p = .030$). Follow-up Bonferroni-adjusted comparisons (resulting in an adjusted α of $p < .025$) revealed that 8- and 9-year-olds did not have significantly higher scores than 6- and 7-year-olds, $t(38) = 1.21$, $p = .24$, $d = 0.392$, while adults had significantly higher scores than 8- and 9-year-olds,

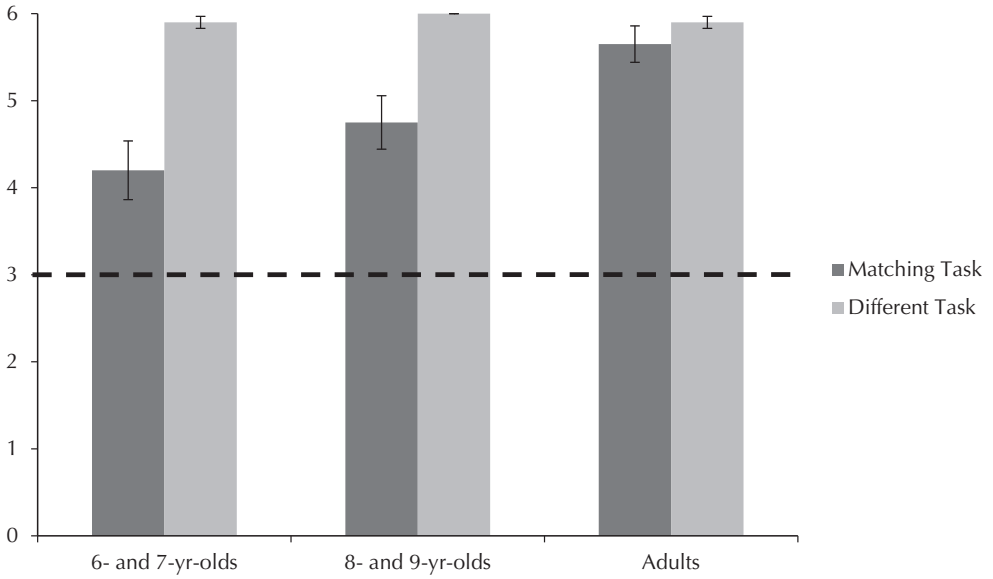

Figure 3. Mean scores in the Conceptual Matching task and Different task in Study 3. For each task, total scores could range from 0 to 6, with a score of 3 (shown with a dotted line) indicating at-chance performance. All age groups' scores significantly exceeded the at-chance score of 3.0 on both tasks. Error bars indicate 1 *SEM* in either direction.

$t(38) = 2.43$, $p = .021$, $d = 0.787$, with significance-level adjustments to account for variance heterogeneity.

Participants from Studies 2 and 3 were recruited from similar populations but at separate times, compromising the validity of cross-study comparisons. With that caveat in mind, an independent samples t test comparing the two older child age groups' Conceptual Matching scores in Studies 2 and 3 found no significant difference between studies, $t(78) = 0.56$, $p = .58$, $d = 0.126$, indicating that the subtler diversity contrasts in Study 3 did not pose problems for participants. Figure 4 shows conceptual matching scores in Studies 2–4.

For the Different task, separate one-sample t tests for each age group, comparing mean scores to the at-chance score of 3.0, found that 6- and 7-year-olds, $t(19) = 42.14$, $p < .001$, $d = 9.422$, 8- and 9-year-olds (no participants scored < 6.0), and adults, $t(19) = 42.14$, $p < .001$, $d = 9.422$, all correctly identified the *diverse machine* significantly more often than would be expected by chance, as shown in Figure 3.

Discussion

Children's strong performance in Study 3 argues against the claim that children are incapable of using subtle types of diversity-based evidence when making judgments of internal complexity. If the primary reason for participants' success in Study 2 was participants' sensitivity to the repeated lexical item in the nondiverse machine's set, then participants would not have succeeded in Study 3. Instead, we have evidence that children as young as 6 and 7 years of age responded to diversity on a deeper level and considered the causal properties of each machine's functions when making their decisions.

Study 4

Study 2 suggested that 4- and 5-year-olds fail to view diversity of function as an indication of underlying complexity. However, the contrasts we used may have been too subtle for young children, as the *diverse machine* differed from the nondiverse machine only in the noun target of one of its functions. In Study 4, we conveyed diversity through verbs as well as nouns. For instance, one *diverse machine* made cupcakes and soups in Study 2, and made cupcakes and wrapped presents in Study 4. If 4- and 5-year-olds' Conceptual Matching task difficulties in Study 2 were solely due to the subtlety of the noun-based contrasts, children of this age should score above chance in Study 4. If 4- and 5-year-olds have an entrenched difficulty using information about diversity to infer underlying complexity, the increased salience of diversity contrasts in Study 4 should not lead to above-chance scores. Because 8- and 9-year-olds scored well in Studies 2 and 3, we did not test children of this age here.

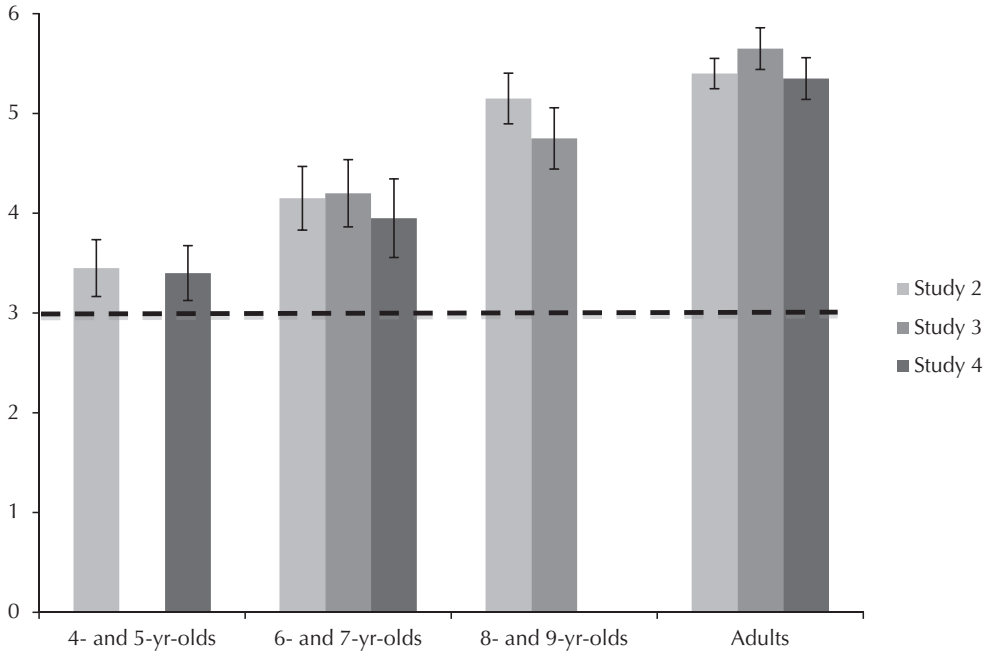

Figure 4. Mean scores in the Conceptual Matching task in Studies 2–4. Total scores could range from 0 to 6, with a score of 3 (shown with a dotted line) indicating at-chance performance. Four- and 5-year-olds were not included in Study 3 and 8- and 9-year-olds were not included in Study 4. Four- and 5-year-olds' scores were not significantly different from the at-chance score of 3.0. All other age groups' scores significantly exceeded chance. Error bars indicate 1 *SEM* in either direction.

Method

Participants

Our final sample of children included twenty 4- and 5-year-olds (15 boys, $M = 61.05$, $SD = 5.65$, range = 52–71) and twenty 6- and 7-year-olds (9 boys, $M = 83.55$, $SD = 6.07$, range = 74–94) tested in our laboratory ($n = 3$), children's museums ($n = 4$), and private schools ($n = 33$). Thirty-three participants were White, four were Black, one was Asian American, one was Latino, and one was biracial. Seven additional participants were excluded due to perseverative responding ($n = 4$), extreme shyness ($n = 1$), severe inattention ($n = 1$), or comprehension difficulties and subsequent failures on the warm-up task ($n = 1$). Our final sample of adults included 20 participants (12 men). Five had a bachelor's degree or higher. Three additional participants were excluded for failing attention and comprehension checks.

Materials and Procedure

The materials and procedure were identical to those in Study 2 with the exception of changes to the *diverse machine*. In Study 4, the *diverse machine* performed a unique action on a unique target noun. The new functions in Study 4 were pretested in a new group of participants, as explained in Appendix S3. The functions of the nondiverse machine and the functions unique to the *diverse machine* received similar technological sophistication ratings from adult participants (all *p*s > .10).

Results

We conducted separate one-sample *t* tests for each age group, comparing mean Conceptual Matching task scores with the at-chance score of 3.0, and found that 4- and 5-year-olds' scores were no different from chance, $t(19) = 1.45$, $p = .16$, $d = 0.325$. Six- and 7-year-olds, $t(19) = 2.41$, $p = .026$, $d = 0.539$, and adults, $t(19) = 11.26$, $p < .001$, $d = 2.518$, matched *complex insides* with the *diverse machines* significantly more often than would be expected by chance, as shown in Figure 4. An ANOVA with Conceptual Matching score as the dependent measure and age as the between-subjects factor revealed a significant main effect for age, $F(2, 57) = 11.05$, $p < .001$, $\eta_p^2 = .279$, whereas Levene's test revealed heterogeneity of variance ($p = .012$). Follow-up Bonferroni-adjusted comparisons (resulting in an adjusted α of $p < .025$) revealed that 6- and 7-year-olds did not have significantly higher scores than 4- and 5-year-olds, $t(38) = 1.15$, $p = .26$, $d = 0.371$, whereas adults had significantly higher scores than 6- and 7-year-olds, $t(38) = 3.14$,

$p = .004$, $d = 1.019$, with significance-level adjustments to account for variance heterogeneity.

For the Different task, we conducted separate one-sample t tests for each age group, comparing mean scores to the at-chance score of 3.0, and found that 4- and 5-year-olds, $t(19) = 17.67$, $p < .001$, $d = 3.950$, 6- and 7-year-olds, $t(19) = 8.79$, $p < .001$, $d = 1.966$, and adults (no participants scored < 6.0) all correctly identified the *diverse machine* significantly more often than would be expected by chance.

General Discussion

In order to succeed at the Conceptual Matching task in Study 1, we believe that participants must know that insides are relevant to a machine's functionality, recognize the visual difference between *simple insides* and *complex insides*, and realize that this difference has implications for functionality, with *complex insides* suggesting broader capabilities than *simple insides*. In order to succeed in Studies 2–4, as did the older children in our sample, participants must also recognize that the different targets have implications for the type of action performed, compare the machines' functions and recognize that one machine has a more diverse set of functions, realize that this difference has implications for the machines' insides, and understand that functional diversity is an indication of underlying complexity.

Our results suggest that 4- and 5-year-old children view number of functions and internal complexity as positively correlated. However, children of this age failed to view an association between diversity of function and internal complexity, even when the diversity contrasts were extreme, as was the case in Study 4. In fact, the increased salience of the diversity contrasts in Study 4 did not lead to higher scores for any age group. If the degree of subtlety with which functional diversity contrasts were conveyed greatly affected participants' performance, then we would have expected substantial shifts in participants' scores across Studies 2–4. Instead, the relative invariance of scores across these studies suggests that the obviousness with which the diversity is conveyed is largely unimportant. The key component for success in our core task is access to the principle that functional diversity is an indication of underlying complexity; those who understand this are likely to do well, and those who lack this understanding are likely to do poorly. We have found a consistent pattern of developmental change across our four studies and shown that children ages 6 and older can make rather sophisticated judgments on the basis of information about diversity. When viewing results from our Conceptual Matching task in conjunction with the Different task, we can infer that what changes with development is not children's ability to detect diversity but rather children's ability to assign meaning to diversity.

Six- and 7-year-olds and 8- and 9-year-olds performed above chance-levels at the Conceptual Matching tasks. Even though young school-age children have failed to consistently assign importance to diversity in animal-based categorical induction tasks, and thus struggle with aspects of scientific reasoning that may be straightforward to adults, we have shown that children of this age succeed at another type of diversity-based scientific reasoning: that of attributing internal complexity to objects that manifest functional diversity. We believe children's strong performance reflects conceptual understanding about the relation between complexity and diversity rather than mere attention to low-level associative features of the task, especially given the younger children's at-chance scores. Imagine that the objects associated with the *diverse machines* were more perceptually interesting than the nondiverse machines' objects (although we believe this is unlikely). If the older children were merely matching the "perceptually interesting" *complex insides* with the machines associated with "perceptually interesting" objects, without understanding the causal significance of the matches, why did the younger children fail to do so? Moreover, children often offered spontaneous explanations that were consistent with genuine conceptual understanding.

The principle that diversity is a cue to internal complexity, which is crucial for success in Studies 2–4, is highly abstract and unlikely to have been formally taught in school. The success of the two older child age groups was apparently due to their increasing ability to access this "diversity implies complexity" principle. Although we do not think 4- and 5-year-olds are wholly incapable of this conceptual understanding (a small number of children from this age scored well in Studies 2 and 4 and produced spontaneous explanations consistent with such an understanding), it seems rare in this age group. The success of 4- and 5-year-olds on Study 1, as well as the results of Erb et al. (2013), argue against an interpretation that children of this age simply fail to match insides to machines on the basis of information about functionality or have

difficulty making judgments about insides more generally.

During the preschool years, children demonstrate an increasing awareness that effects arise from causes, which can often exert their influence through underlying mechanisms (Buchanan & Sobel, 2011; Erb et al., 2013). What factors may give rise to the understanding that diversity is a cue to complexity, and why does that understanding take longer to emerge than linking number of functions to complexity? One reason may simply be greater experience with different degrees of diverse behaviors within distinct categories. With age, children likely gain increasing exposure to artifacts and animals with different ranges of actions, functions, and behaviors, and such exposure may highlight the presence of diversity, which children may seek to explain through underlying causes—that is, confronted with more information about variation of diversity within a category, children may seek out a reason. Direct experience with causal mechanistic interventions, such as assembling a mechanical toy car or operating a circuit board, may not be necessary for this understanding but is likely to speed its emergence (see Sheridan et al., 2014, for an account of how experience with circuit boards can focus children's attention on mechanistic information). Given that even preschoolers demonstrate highly abstract knowledge regarding the kinds of "hidden insides" that animals and artifacts have (Gottfried & Gelman, 2005), despite rarely witnessing them first hand, it is not implausible for young children to form understandings about the complexity of those insides.

Once children grasp the link between diversity and underlying complexity, we believe they may engage with the objects around them in qualitatively different ways. Much as causally confounded objects motivate increased exploratory play in preschoolers (Schulz & Bonawitz, 2007), diverse or variable objects may motivate information-seeking searches and questions, as the objects' range of actions indicate causal complexity and warrant further explanation. Our findings suggest that, as children grow older, diversity takes on increased meaning as a signal of causal significance.

Although our method differs from other studies of diversity-based reasoning on many dimensions, our findings of strong performance in young school-age children may shed light on the failures of children to act on information about diversity in many other inductive reasoning tasks before reaching the age of 9. According to Rhodes and Liebenson (2015), children's preferences for typical exemplars within familiar categories, especially animals, interfere with children's diversity-based reasoning in the domain of biology. In the domain of machines, children may lack expectations for "typical exemplars." Moreover, the artifacts we used in our study were novel. Thus, our task reduced some of the challenges posed by studies using real-world animal categories. Young children may not always grasp the causal significance of internal or "hidden" properties (see also Ahn, Gelman, Amsterlaw, Hohenstein, & Kalish, 2000; Heit & Hahn, 2001). Categorical induction paradigms often introduce participants to unfamiliar internal biological features, such as "a green spot in [the] mouth" (Gutheil & Gelman, 1997), that do not seem "causally central," and young children may struggle to draw inferences from such features. One reason why young children succeeded in our study may be that children correctly understood that a machine's insides have a role in regulating its functions and are "causally central," even though our script lacked explicit instruction to this effect (see Hadjichristidis, Sloman, Stevenson, & Over, 2004; Keil et al., 1998, on feature and causal centrality). Diversity-based induction studies may find successes in younger children if the significance of the properties in question is made salient, that is, stating that a given property causes or enables a specific goal. Finally, in our study, diversity was manifested within a single object, through its functions, rather than across several objects, as is the case in most diversity-based reasoning studies. It may be more computationally tractable for children to reason about diversity when presented within a single entity rather than across a sample of several entities.

Given that our machine examples and insides images were not drawn directly from real-world machines, one could argue that there is no "right" way to complete the Conceptual Matching task. One could also claim that technological advances often involve a process of paring down extraneous features, such that technological sophistication or functional diversity may not indicate internal complexity. However, we believe the "complexity enables diversity" principle can be considered normative. In our studies, few participants of any age seemed to think that "simplicity enables diversity," and almost all adults' scores indicated a "complexity enables diversity" belief. Participants' spontaneous explanations for their choices revealed that, compared to *simple insides*, most participants viewed *complex insides* as having a more complex overall mechanistic structure or multiple distinct

mechanisms in place. No participants stated that *complex insides* seemed obsolete or inefficient.

Although the adult-like pattern in our study was to privilege information about diversity when making complexity judgments, diversity was not the only cue that participants could have used (for exceptions to the diversity principle in adults' reasoning, see Proffitt, Coley, & Medin, 2000). For instance, participants may have matched *complex insides* with whichever machine performed a function they deemed most optimal. Such a strategy could occasionally have caused participants to match *complex insides* with the nondiverse machine. It is possible that even the youngest participants in our sample viewed diversity as an indication of complexity, but diversity was just one of many cues that such children used. According to this account, the developmental shift driving our findings was not the emergence of the diversity principle per se but rather a shift in the weights assigned to different complexity cues, with diversity taking on increased significance with age and overriding other cues.

Most of our child participants were White and came from middle-class households in the Northeast region of the United States, raising concerns about the generalizability of our findings. We have no reason to believe children's race influenced their performance, although their SES may have done so. Middle- and high-SES parents may provide exposure to enriched environments, such as museums and "makerspaces," that allow for encounters with the types of technological and mechanical devices that may give rise to insights about the connections between functional diversity and underlying complexity, as well as facilitate structured conversations about such devices (Benjamin, Haden, & Wilkerson, 2010; Jant, Haden, Uttal, & Babcock, 2014; Sheridan et al., 2014). We believe that general patterns of increasing performance on the Conceptual Matching task with age, and simultaneous Conceptual Matching task failure and Different task success in younger age groups, would hold true in many populations, although the exact ages at which success is achieved could be population specific and influenced by SES.

Our study focused on complexity in the domain of human-made artifacts. Individuals living in nonindustrialized cultures could show lower rates of success at our machine-based tasks, although all cultures do have tools of varying complexity and multiple functions. However, even if members of such traditional cultures show less sensitivity to the role of diversity as a cue to artifact complexity, they may nonetheless demonstrate sensitivity to diversity as an indication of underlying complexity in the domain of biology. Further studies can explore children's and adults' sensitivity to diversity, plus elements such as number and speed of functions, as indications of complexity in a variety of domains. Further studies can also test the visual and spatial cues individuals use to detect the presence of complexity, such as the number, type, and interconnectedness of parts (see Erb et al., 2013; Gelman, 1988). Finally, school- or museum-based interventions, including first-hand experiences with the internal mechanisms of machines, can be designed and tested to determine if they improve children's performance on tasks such as those described here (Benjamin et al., 2010; Sheridan et al., 2014).

Very young children, and even infants, make inferences about unseen causal mechanisms from surface properties, suggesting that the propensity to discover hidden factors and causes is a foundational operating principle of the human mind. There are many dimensions of surface variation that can serve as cues to internal operations and some, such as number of surface functions, may be linked to complexity inferences quite early in development. Other dimensions, such as diversity, although salient in their own right, do not drive strong inferences about insides until the school years. It may take many more years, and ultimately the assistance of formal sciences, to be able to optimally infer the unseen from the seen.

References

Ahn, W. K., Gelman, S. A., Amsterlaw, J. A., Hohenstein, J., & Kalish, C. W. (2000). Causal status effect in children's categorization. *Cognition*, *76*, B35–B43. doi:10.1016/S0010-0277(00)00077-9

Benjamin, N., Haden, C. A., & Wilkerson, E. (2010). Enhancing building, conversation, and learning through caregiver–child interactions in a children's museum. *Developmental Psychology*, *46*, 502. doi:10.1037/a0017822

Buchanan, D. W., & Sobel, D. M. (2011). Mechanism-based causal reasoning in young children. *Child Development*, *82*, 2053–2066. doi:10.1111/j.1467-8624.2011.01646.x

Erb, C. D., Buchanan, D. W., & Sobel, D. M. (2013). Children's developing understanding of the relation between variable causal efficacy and mechanistic complexity. *Cognition*, *129*, 494–500. doi:10.1016/j.cognition.2013.08.00

Gelman, S. A. (1988). The development of induction within natural kind and artifact categories. *Cognitive Psychology*, *20*, 65–95. doi:10.1016/0010-0285(88)90025-4

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition*, *38*, 213–244. doi:10.1016/0010-0277(91)90007-Q

Gottfried, G. M., & Gelman, S. A. (2005). Developing domain-specific causal-explanatory frameworks: The role of insides and immanence. *Cognitive Development*, *20*, 137–158. doi:10.1016/j.cogdev.2004.07.003

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

Hadjichristidis, C., Sloman, S., Stevenson, R., & Over, D. (2004). Feature centrality and property induction. *Cognitive Science*, *28*, 45–74. doi:10.1207/s15516709cog2801_2

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

Heit, E., Hahn, U., & Feeney, A. (2004). 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: APA.

Jant, E. A., Haden, C. A., Uttal, D. H., & Babcock, E. (2014). Conversation and object manipulation influence children's learning in a museum. *Child Development*, *85*, 2029–2045. doi:10.1111/cdev.12252

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

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences*, *7*, 368–373. doi:10.1016/S1364-6613(03)00158-X

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition*, *65*, 103–135. doi:10.1016/S0010-0277(97)00041-3

Kim, N. S., & Keil, F. C. (2003). From symptoms to causes: Diversity effects in diagnostic reasoning. *Memory & Cognition*, *31*, 155–165. doi:10.3758/BF03196090

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

Newman, G. E., Herrmann, P., Wynn, K., & Keil, F. C. (2008). Biases towards internal features in infants' reasoning about objects. *Cognition*, *107*, 420–432. doi:10.1016/j.cognition.2007.10.006

Noyes, A., & Christie, S. (in press). Children prefer diverse samples for inductive reasoning in the social domain. *Child Development*. Advance online publication. doi:10.1111/cdev.12522

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

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

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 and Development*, *9*, 112–143. doi:10.1080/15248370701836626

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

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

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*, 1045–1050. doi:10.1037/0012-1649.43.4.1045

Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 15937–15942. doi:10.1073/pnas.1314075110

Sheridan, K., Halverson, E. R., Litts, B., Brahms, L., Jacobs-Priebe, L., & Owens, T. (2014). Learning in the making: A comparative case study of three makerspaces. *Harvard Educational Review*, *84*, 505–531. doi:10.17763/haer.84.4.brr34733723j648u

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

Simons, D. J., & Keil, F. C. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition*, *56*, 129–163. doi:10.1016/0010-0277(94)00660-D

Sobel, D. M., Yoachim, C. M., Gopnik, A., Meltzoff, A. N., & Blumenthal, E. J. (2007). The blicket within: Preschoolers' inferences about insides and causes. *Journal of Cognition and Development*, *8*, 159–182. doi:10.1080/15248370701202356

Appendix A
Text for Conceptual Matching Task Item #3 in Studies 1–4

Study 1

This machine can make cupcakes.

This machine can make cupcakes and soups.

Study 2

This machine can make cupcakes that look like this. It can also make cupcakes that look like this.

This machine can make cupcakes that look like this. It can also make soups that look like this.

Study 3

This machine can make cupcakes that look like this. It can also make muffins that look like this.

This machine can make cupcakes that look like this. It can also make soups that look like this.

Study 4

This machine can make cupcakes that look like this. It can also make cupcakes that look like this.

This machine can make cupcakes that look like this. It can also wrap presents that look like this.

Note. The prompts were read aloud by the experimenter for child participants and presented via text for adult participants. For the sake of simplicity, the *one-function* and nondiverse machines are listed first, although the order was randomized for participants. To facilitate the participants' process of comparison, the targets and functions associated with both machines (in this case, the cupcake, which was brown in the accompanying photograph) were always displayed and discussed first for each machine.

Appendix B
Study Flow for Studies 1–4

Study 1

Introduction
Warm-up
Conceptual Matching task
Attention check

Studies 2–4

Introduction
Warm-up
Conceptual Matching task
Attention check
Different Introduction task
Different task

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's website:

Appendix S1. Complete Stimuli for Conceptual Matching Task in Studies 1–4.

Appendix S2. Percentage of Participants Who Matched Complex Insides to the Two-Function Machine or Diverse Machine for Each Age Group and Item.

Appendix S3. Explanation of Pretesting Procedure for Adults' Ratings of Individual Functions.

Appendix S4. Different Introduction Task Items.