Running head: Evidential diversity and premise probability

Evidential diversity and premise probability in young children's inductive judgment

Yafen Lo
Rice University
ph: 713-348-5831
fx: 713-348-5221
yafen@rice.edu

Ashley Sides
Rice University
ph: 713-348-5831
fx: 713-348-5221
sides@rice.edu

Joseph Rozelle
Rice University
ph: 713-348-5831
fx: 713-348-5221
jarozelle@home.com

Daniel Osherson
Rice University
ph: 713-348-5831
fx: 713-348-5221
osherson@rice.edu

September 1, 2001

Keywords: Psychology, Cognitive Development, Human Experimentation.

Abstract

A familiar adage in the philosophy of science is that general hypotheses are better supported by varied evidence than by uniform evidence. Several studies suggest that young children do not respect this principle, and thus suffer from a defect in their inductive methodology. We argue that the diversity principle does not have the normative status that psychologists attribute to it, and should be replaced by a simple rule of probability. We then report experiments designed to detect conformity to the latter rule in children's inductive judgment. The results suggest that young children in both the United States and Taiwan are sensitive to the constraints imposed by the rule on judgments of probability and evidential strength. We conclude with a suggested reinterpretation of the thesis that children's inductive methodology qualifies them as "little scientists."

Evidential diversity and premise probability in young children's inductive judgment

1 Introduction

A central issue in cognitive development is whether children's scientific reasoning is methodologically sound (simply short on facts), or else neglectful of fundamental principles of inductive reasoning (Carey, 1985; Gopnik & Meltzoff, 1996; Keil, 1989; Koslowski, 1996; Kuhn, 1989; Markman, 1989). To address the issue, normative standards of inductive reasoning must be formulated, and children's thinking evaluated in their light. One candidate principle concerns the diversity of evidence in support of a given hypothesis. It is widely claimed that greater diversity entails greater support (e.g., Franklin & Howson, 1984; Hempel, 1966). Several experiments suggest that young children fail to perceive this relation (Carey, 1985; Gutheil & Gelman, 1997; López, Gelman, Gutheil & Smith, 1992). Consider, for example, arguments (a) and (b), below. The premises of the former provide more diverse evidence than do the premises of the latter. Yet the five-to-six year olds described in López et al. (1992) gave no sign of perceiving greater support for the conclusion of (a) compared to (b).

(a)
Cats have leukocytes inside.
Buffalo have leukocytes inside.

All animals have leukocytes inside.

(b)
Cows have leukocytes inside.
Buffalo have leukocytes inside.

All animals have leukocytes inside.

It is thus tempting to conclude that young children's inductive methodology is faulty.

In contrast to children, American college students often rely on diversity, as documented in Osherson, Smith, Wilkie, López and Shafir (1990), and in the elegant study by López (1995). [Diversity based reasoning in other cultures is discussed in Atran (1995), Choi, Nisbett & Smith (1997), López, Atran, Coley, Medin & Smith (1997), and Viale & Osherson (2000).] Such developmental differences in the use of diversity seem to discredit the picture of children as "little scientists" since bigger scientists exploit a principle of reasoning that is unavailable at an earlier age.

The present discussion is devoted to arguments whose conclusions imply their premises [as in (a), (b), above]. Such arguments are called "general." We attempt to ascertain whether children's reasoning about such arguments is deficient from the normative point of view. The issue is therefore not merely whether children can detect evidential diversity. An ingenious study by Heit and Hahn (1999) reveals circumstances in which five and six-year-olds prefer diverse to homogeneous evidence when evaluating hypotheses. For example, they find argument (c), below, to be stronger than (d).

(c)
Jane likes strawberry ice cream.
Jane likes vanilla ice cream.
Jane likes pistachio ice cream.

Jane likes blackcurrant ice cream.

(d)
Jane likes chocolate ice cream.

Jane likes blackcurrant ice cream.

This finding leaves open, however, whether young children use information about diversity in the evaluation of general arguments. The arguments above are not general, and it is possible that the preference for (c) rests partly on similarity (for example, the similarity of strawberry to blackcurrant ice cream, both being fruit flavors). To limit attention to general arguments, "specific" conclusions [like those in (c) and (d)] are set aside.

We proceed as follows. The next section questions the use of the diversity principle as a normative standard of reasoning. In its place we propose a probabilistic principle whose presence in young children's thinking has not been assessed heretofore. It is then shown that young children are sensitive to the latter principle, especially when arguments are evaluated in the context of a detective game. Finally, we discuss the bearing of these results on the "little scientist" picture of cognitive development.

2 Normative Considerations

The present section bears on normative (or "prescriptive") issues rather than empirical fact. The goal is to articulate a sound principle of scientific inference to which young children's reasoning may be compared. Only in the light of such principles can the quality of children's reasoning be meaningfully assessed. Analogously, the theory of arithmetic is prerequisite to tracing the development of arithmetical competence. The normative theory of scientific inference enjoys less consensus than arithmetic, however, and it can be

challenging to formulate sound prescriptions for reasoning. We now attempt to formulate one such prescription, beginning with critical evaluation of the diversity principle. No empirical assertions about actual reasoning are made in this section; only the criterion of correct reasoning is at issue.

2.1 The diversity principle and scientific reasoning

For use as a standard against which to evaluate intuitive reasoning, psychologists typically restrict the diversity principle to statements of subject-predicate form, where the subject is a natural category (like *lion*) and the predicate is "blank." Blank predicates are indefinite in their application to given categories, but clear enough to communicate the kind of property in question. For example, *has an ulnar artery* is blank since most reasoners cannot determine the animals to which it applies, yet it clearly expresses a biological property. (The earliest use of blank predicates to study inductive reasoning appears to be Rips, 1975.)

Given a category C and a blank predicate P, we denote by (C, P) the assertion that P applies to the members of C (for example, that lions have ulnar arteries). By an *argument* is meant a finite list of such assertions. The last statement on the list is called the argument's *conclusion*, the remaining statements are its *premises*. Arguments are displayed vertically as in (a) - (d) above, or else horizontally with the conclusion set off by a slash. An argument is *strong* to the extent that the premises give reason to believe the conclusion. The preceding terminology is familiar from inductive logic (see Gustason, 1994), and from recent studies of inductive judgment (e.g., Choi et al., 1997).

For simplicity we restrict attention to arguments of the form $(X_1, P), (X_2, P)$ / (Y, P), in which the same blank predicate P appears in all statements, and the conclusion category Y includes the premise categories X_1, X_2. Along with (a) and (b) above, two such argument are:

(e)
Hippopotamuses have an ulnar artery.
Rhinoceroses have an ulnar artery.

All mammals have ulnar arteries.

(f)
Hippopotamuses have an ulnar artery.
Hamsters have an ulnar artery.

All mammals have ulnar arteries.

The diversity principle is meant to apply at the normative level to pairs of arguments such as the foregoing, ruling (f) to be the stronger of the two. Thus, Heit and Hahn (1999) refer

to the "strong case to be made for the normative status of diversity-based reasoning," and López (1995, p. 374) claims that "to test whether all mammals have an ulnar artery, you should examine [along with hippos] a hamster instead of a rhinoceros because hippopotamuses and hamsters are more diverse than hippopotamuses and rhinoceroses." As these passages indicate, judgments contrary to the diversity principle are sometimes considered to be in error.

2.2 Objections to the diversity principle

The diversity variable, however, is not related in such simple fashion to argument strength, even at the prescriptive level. This is because so-called "blank" predicates often retain enough meaning to open the door to legitimate reasoning that violates the diversity principle. An example is provided by the predicate *often carry the parasite Floxum*. It counts as blank since the parasite *Floxum* is unfamiliar, so nothing can be divined about which mammals are more or less likely to suffer from it. Now consider the following arguments.

(g)
Housecats often carry the parasite Floxum.
Fieldmice often carry the parasite Floxum.

All mammals often carry the parasite Floxum.

(h)
Housecats often carry the parasite Floxum.
Tigers often carry the parasite Floxum.

All mammals often carry the parasite Floxum.

It seems undeniable that housecats resemble tigers more than they resemble fieldmice. Yet it appears perfectly defensible to judge argument (h) to be stronger than (g), on the grounds that housecats might catch *Floxum* from fieldmice (their prey) whereas they have little contact with tigers.[1] We are not claiming that everyone (or even a majority of people) will think of the predator-prey reason for judging (h) to be stronger than (g). Rather, we claim that it is perfectly *reasonable* to make such a judgment (because of the predator-prey relation or on some other basis). In this case, reasonable judgments about strength run counter to the advice offered by the diversity principle. Hence, the latter is discredited

as a basis for evaluating inductive intuition. Of course, the example does not show that diversity *always* gives the wrong advice, only that it *sometimes* does. But this is enough to undermine its claim to normative status.

Perhaps it will be objected that the similarity of categories must be evaluated in light of the predicate in play. When it comes to parasites, housecats resemble fieldmice more than tigers, so the diversity principle gives the right advice after all. One difficulty with this suggestion is that it makes application of the diversity principle dependent on prior inductive inference. Before making an assessment of diversity, we must engage in substantive scientific reasoning, namely, about how similarity is to be assessed in the context of a given predicate. For example, it must be determined that animals in a predatory relation are similar when the predicate concerns parasites but not (say) color. Such determination is a scientific affair that already requires sound judgment about argument strength.[2] More generally, diversity seems to be of no greater help in evaluating hypotheses than our hypotheses are in perceiving diversity. Thus, Earman (1992, p. 79) reminds us that "before the scientific revolution the motions of the celestial bodies seemed to belong to a different variety than the motions of terrestrial projectiles, whereas after Newton they seem like peas in a pod."[3]

In any event, it is easy to produce violations of the diversity principle even when similarity is assessed relative to a predicate. Consider *can scratch through Bortex fabric in less than 10 seconds*. The unfamiliarity of Bortex makes it impossible to tell which animals can scratch through it, although it allows sensible inferences about the conditional probability of one animal scratching through assuming that another does. In this respect, the Bortex predicate is as blank as others that appear in the literature, e.g., *has an ulnar artery*. (Observe that the conditional probability of wolves having an ulnar artery given that coyotes do is sure to be judged higher than the conditional probability of wolves having such an artery given that moles do.) Relative to the Bortex predicate, squirrels and mice are more similar to each other than are squirrels and bears. Nonetheless, it is clear that the homogeneous premises of argument (j), below, convey greater strength than the diverse premises of argument (i), in contradiction with the diversity principle.

(i)
Squirrels can scratch through Bortex fabric in less than 10 seconds.
Bears can scratch through Bortex fabric in less than 10 seconds.

All forest mammals can scratch through Bortex fabric in less than 10 seconds.

(j)
Squirrels can scratch through Bortex fabric in less than 10 seconds.
Mice can scratch through Bortex fabric in less than 10 seconds.

All forest mammals can scratch through Bortex fabric in less than 10 seconds.

The only way to avoid these kinds of examples is to restrict the diversity principle to statements whose predicates are entirely empty, as in "Lions have property P." In this case, however, the discussion will have strayed so far from genuine scientific reasoning as to become irrelevant to questions about inductive methodology. Research on category-based reasoning has therefore avoided nonce predicates like *has P* (see, e.g., Osherson et al., 1990; Rips, 1975). The better solution is to formulate principles of inductive strength that are not restricted to meaningless predicates. The principle proposed below meets this condition.[4]

We conclude that the diversity principle should not be used as a standard against which to assess the methodological soundness of children's scientific reasoning. Diverse premises often confer inductive strength, but exceptions to this rule suggest the presence of an underlying variable that mediates the impact of evidential diversity. We now attempt to identify this variable.

2.3 A probability principle

There is little doubt that the premises appearing in (f) confirm *all mammals have an ulnar artery* more than do the premises in (e). If diversity is not the operating variable in this difference, what is? One response comes from Bayesian probability theory interpreted as a normative account of scientific inference. The latter theory provides a systematic explanation of much scientific practice (Earman, 1992; Howson & Urbach, 1993), and has figured prominently in psychologists' assessment of adult reasoning (Kahneman & Tversky 1996). There are certainly competing philosophies of science (e.g., Martin & Osherson, 1998; Mayo, 1996), and not all psychologists are comfortable with the Bayesian picture (Gigerenzer, 1991, 1996). But Bayesianism has the virtue of being explicit, and it offers a simple explanation for the examples discussed above.

Following an analysis advanced by Horwich (1982), we propose that in the kinds of arguments at issue here, it is the probability of the premises prior to accepting the conclusion that governs strength. To get the analysis off the ground, it is necessary to give a probabilistic interpretation of argument strength. Recall that an argument is strong to the extent that its premises give reason to believe its conclusion. Thus in a strong argument, the probability of the conclusion given the premises should be greater than the probability of the conclusion alone. An argument with premises A, B and conclusion C should therefore be strong to the extent that $\mathrm{P}(C \mid A \wedge B) > \mathrm{P}(C)$.[5] There are many ways to quantify the extent to which $\mathrm{P}(C \mid A \wedge B)$ exceeds $\mathrm{P}(C)$. Following other authors, like Myrvold (1996), we rely on the ratio of the conditional to the unconditional probability. Officially:

> Definition: Let argument $A, B \mathbin{/} C$ be given. Its strength is defined to be $\dfrac{\mathrm{P}(C \mid A \wedge B)}{\mathrm{P}(C)}$.

Premises confirming a conclusion thus yield strength greater than 1, and premises infirming a conclusion yield strength less than 1.

The arguments figuring in the present discussion are all general, and therefore have the particularity that the conclusion logically implies each premise (since the same predicate occurs in all statements, and the conclusion category includes the premise categories). Under these conditions, argument strength stands in a simple relation to premise-probability.

> Premise probability principle (*PPP*): Suppose that arguments $A, B \mathbin{/} C$ and $A', B' \mathbin{/} C$ are given, where C logically implies A, A', B, and B'. Then the strength of the first argument is greater than the strength of the second if and only if $\mathrm{P}(A \wedge B) < \mathrm{P}(A' \wedge B')$.

> *Proof:* By definition, the strength of the first argument is greater than the strength of the second if and only if
>
> $$\frac{\mathrm{P}(C \mid A \wedge B)}{\mathrm{P}(C)} > \frac{\mathrm{P}(C \mid A' \wedge B')}{\mathrm{P}(C)}.$$
>
> By familiar properties of conditional probability, the latter inequality holds if and only if
>
> $$\frac{\mathrm{P}(C \wedge A \wedge B)}{\mathrm{P}(C) \times \mathrm{P}(A \wedge B)} > \frac{\mathrm{P}(C \wedge A' \wedge B')}{\mathrm{P}(C) \times \mathrm{P}(A' \wedge B')}.$$

Since C implies A, A', B, and B', the last inequality holds if and only if

$$\frac{\mathrm{P}(C)}{\mathrm{P}(C) \times \mathrm{P}(A \wedge B)} > \frac{\mathrm{P}(C)}{\mathrm{P}(C) \times \mathrm{P}(A' \wedge B')},$$

which holds if and only if

$$\frac{1}{\mathrm{P}(A \wedge B)} > \frac{1}{\mathrm{P}(A' \wedge B')},$$

thus if and only if

$$\mathrm{P}(A \wedge B) < \mathrm{P}(A' \wedge B'),$$

which proves the claim.

The upshot is that the less probable the premises, the stronger the argument. Diversity of premises is often (but not systematically) associated with low premise probability. For example, it seems less likely that Hippos and Hamsters have ulnar arteries than that Hippos and Rhinos do. Based on this assumption, *PPP* accounts for the greater strength of (f) compared to (e). Likewise, given their unrelated habitats, housecats and tigers seem less likely to carry common parasites than do housecats and mice. For anyone sharing the latter judgment, *PPP* rules (h) to be stronger than (g). The relative probability of the premises in (i) and (j) similarly make it reasonable to judge the latter to be stronger than the former.[6]

From the proof of *PPP* it can be seen that the principle applies to arguments with any predicate, blank or not. Its normative status is thereby enhanced since inductive reasoning almost invariably involves meaningful predicates. Observe also that *PPP* is a consequence of no more than the axioms of probability along with our definition of argument strength (which could be replaced by many alternatives without altering *PPP*). Its normative appeal thus derives from the theory of probability, which is independently motivated (see Osherson, 1995; Skyrms, 1986, for discussion of the justification of the probability axioms). Note that we have formulated *PPP* in terms of arguments with two premises. But the same principle holds for any number of premises, including just one. For simplicity in what follows, however, we rely on two-premise arguments with blank predicates to test children's respect for *PPP*.

2.4 Assessing conformity to *PPP*

It is important to recognize that *PPP* imposes no specific assessments of argument strength; it merely relates strength to the probability of an argument's premises. Thus, it is no sign of methodological weakness if children do not find (a) to be stronger than (b). They may simply take the premises of the former to be more likely than the premises of the latter (thus believing it to be more likely that cats and buffalo have leukocytes inside than that cows and buffalo do). However ill-informed such a judgment might appear to adults, it violates no law of proper scientific methodology. Indeed, children might reasonably suppose that cows have only milk inside, which leads directly to low probability for the premises of (b), hence to high strength for (b). On the other hand, it *is* incompatible with *PPP* for children to simultaneously believe either

- that the premises of (a) are less likely than those of (b) *and* that (b) is stronger than (a), or
- that the premises of (b) are less likely than those of (a) *and* that (a) is stronger than (b).

More generally, consider two general arguments X_1, X_2 / Z and Y_1, Y_2 / Z with the same conclusion Z. [Arguments (a), (b) illustrate such a pair.] For these arguments, a reasoner *respects PPP* if either (i) she believes that $X_1 \wedge X_2$ is less likely than $Y_1 \wedge Y_2$ and that the first argument is stronger than the second, or (ii) she believes that $Y_1 \wedge Y_2$ is less likely than $X_1 \wedge X_2$ and that the second argument is stronger than the first. Thus, respect for *PPP* requires no specific opinion about premise-probability or about argument strength; only the relation between the opinions is at issue. In particular, assessing respect for *PPP* has no connection to the relative difficulty of the two kinds of judgments, nor with how often the reasoner gets either "right" (which has no meaning within the current perspective). Since relative difficulty of the two kinds of judgment is not at issue, it makes no difference to assessing respect for *PPP* whether the format of premise-probability questions is different from the format of argument-strength questions. Each question should be posed in whatever manner maximizes its clarity to the reasoner. Respect for *PPP* may then be evaluated by comparing answers to the two types of question, as described in (i), (ii). Such is the methodology of the experiments reported below.[7]

Using college students as subjects, López (1995) reports a strong negative correlation between premise-probability and an indirect measure of argument strength; similar findings are reported below. College students appear therefore to respect *PPP*. In contrast, no test of young children's conformity to *PPP* has yet been carried out. Such tests are now reported.

3 Experiments

All of our experiments were designed to collect judgments about argument strength versus premise probability in two-premise general arguments with blank predicates. The crucial items were arguments evaluated at separate times for strength and premise probability. In what follows, the crucial questions involving strength will be called PPP_{stren} items. The crucial questions involving premise probability will be called PPP_{prob} items. As noted earlier, advocates of the diversity principle claim that sound methodology requires children to respond to PPP_{stren} items as adults do (namely, in terms of diversity). In contrast, *PPP* requires only that responses to the PPP_{stren} items be concordant with responses to the PPP_{prob} items, in the sense discussed above. Specifically, arguments with less probable premises should be judged to be stronger.

In all of the experiments, the PPP_{stren} and PPP_{prob} items were mixed with other questions relevant to inductive phenomena known as "monotonicity" and "typicality" in the previous literature (see Osherson et al. 1990). Monotonicity concerns the choice between arguments of forms $A, B/C$ and B/C, where C is a general conclusion that implies both A and B. Judgment is called "monotonic" if the first argument of the pair is deemed stronger. In fact, such a choice can be conceived as an application of *PPP* inasmuch as the claims A, B are less likely to both be true than is the constituent claim B. Typicality concerns the choice between arguments of forms A/C and B/C, where C is a general conclusion that implies both A and B. Judgment conforms to the typicality phenomenon if the stronger argument is taken to be the one with more typical premise category relative to the category of C.[8] The presence of monotonicity and typicality items allowed us to verify that children understood the nature of questions about argument strength. In fact, as seen below, our children tended to respond like adults to these items. All the arguments figuring in the experiments relied on predicates designed to be blank to adult reasoners.

Experiments 1 and 2 involved young children, and shared some features that can be mentioned here. To communicate effectively, we relied on two fictitious characters, the

detectives Max and Morgan. The detectives were trying to learn facts about animals, fruit, etc., and the child was to help in their investigation. Questions involving premise-probability (namely, the PPP_{prob} items) were phrased in terms of surprise. Specifically, each detective learned the facts comprising the premises of one of the two arguments to be compared. The children were asked which detective was surprised (or alternatively, not surprised) by his discovery. We interpreted the corresponding set of premises as having lesser (alternatively, greater) probability. Questions involving strength (namely, all other items, including PPP_{stren}) were phrased in terms of clues. Each detective learned the facts comprising the premises of one of the two arguments, and the children were asked who had the better clue for the argument's conclusion.

In all cases, children were pre-tested to make sure they were familiar with the animals about which they would be questioned. Specifically, for each animal, the children were asked to name its picture, and indicate its approximate size. Although corrections during the pretest were occasionally necessary, no subject was dropped on this basis.

3.1 Experiment 1

Participants. Preschool children served as subjects. They were recruited from day care centers in a middle class neighborhood of Houston, Texas. The sample included 18 boys and 25 girls (age range = 2.92 to 5.92 years; mean age = 4.53 ; SD = .91). Two children were excluded from the analysis due to absence from the second interview session.

Design. Each child was exposed to (a) four arguments testing his/her response to typicality, (b) four arguments testing his/her response to monotonicity, (c) four arguments assessing his/her intuitions about premise probability (the PPP_{prob} items), and (d) four arguments testing the strength of arguments involving the same premises as in (c) (the PPP_{stren} arguments). Thus, a total of 16 questions were posed to each child. See Appendix A for the stimuli used.

The sixteen questions were randomly distributed over two testing sessions (on different days) with the constraint that neither of the sessions for any child include both a PPP_{prob} and PPP_{stren} question relative to the same argument. (So, children never saw the same premises twice in a given session.)

Materials. Laminated color photographs were used as visual support for each category mentioned in the premise of an argument. No pictures were used for superordinate categories like *birds, animals,* and *fruit.*

Procedure. The experiment was conducted in two 15-minute sessions with at least one day between. Participants were tested individually. Children were told that two detectives (Max and Morgan) had found some clues about different animals/fruits/vegetables. The clues were category-predicate sentences, with pictures of the relevant categories serving as visual support. We illustrate the instructions with one PPP_{prob} item and its corresponding PPP_{stren} item.

> **Instructions for a PPP_{prob} item:** Detectives Max and Morgan found some clues. After they saw their clues, one of them said "I knew it would be like that. I am not surprised at all!" But we don't know who said it. Max found that lions and tigers have thick blood. Morgan found that lions and rhinos have thick blood. Which one of them was not surprised? Do you think Max was not surprised because he found that lions and tigers have thick blood, or Morgan was not surprised because he found that lions and rhinos have thick blood? (The order of presentation of the two sets of clues was individually randomized.)

> **Instructions for a PPP_{stren} item:** Detectives Max and Morgan wanted to know if all animals have thick blood. Both of them found some clues. Max found that lions and tigers have thick blood. Morgan found that lions and rhinos have thick blood. If they wanted to know whether *all animals* have thick blood, which one is the better clue? Is this the better clue, that lions and tigers have thick blood? Or is this the better clue, that lions and rhinos have thick blood? (Again, the order of presentation of the two sets of clues was individually randomized.)

Notice that conformity to *PPP* required the child to choose different clues across corresponding PPP_{prob} and PPP_{stren} items. For example, they would need to choose lions and tigers in the first item and lions and rhinos in the second. In this way we sought to reduce spurious success due to response perseverance.

Instructions for the monotonicity and typicality questions followed the same format as the PPP_{stren} item (indeed, all three types of items involve argument strength).

Results. For each child, we now define five scores. The scores are called *typicality, monotonicity, premise-probability, diversity,* and *PPP*. They each range from 0 to 4. In each case, a score of 2 would be expected by chance responding.

For typicality, we assigned the child one point for each typicality item in which s/he judged the argument with more typical premise-category to be stronger. For monotonicity, we assigned the child one point for each monotonicity item in which s/he judged the argument with two premises to be stronger. For premise-probability, we assigned the child one point for each PPP_{prob} item in which s/he judged the more homogeneous premise pair to be more probable. (Homogeneity was assessed in the obvious biological sense.) For diversity, we assigned the child one point for each item in which s/he judged the argument with more diverse premises to be stronger. For *PPP*, we compared the child's answers to the PPP_{prob} questions and the PPP_{stren} items. Consider a choice between the premise pairs, P_1, P_2 versus P_1, P_2', along with the corresponding choice between the PPP_{stren} arguments $P_1, P_2/C$ versus $P_1, P_2'/C$. We assigned the child one point whenever s/he claimed either that:

- the premise pair P_1, P_2 was more probable than the premise pair P_1, P_2', and the argument $P_1, P_2'/C$ was stronger than the argument $P_1, P_2/C$; or
- the premise pair P_1, P_2' was more probable than the premise pair P_1, P_2, and the argument $P_1, P_2/C$ was stronger than the argument $P_1, P_2'/C$.

Let it be stressed that no normative significance is attached to the typicality, premise-probability or diversity scores. We compute these scores only in the interest of comparing children's reasoning to that of adults (and because these scores have appeared in the previous literature devoted to the diversity principle). As observed above, the monotonicity score does reflect a principle of sound inference (at least, from the Bayesian point of view), but the choice between arguments in these questions seems obvious, even for children. It is only the *PPP* score that potentially carries new information about the quality of children's scientific reasoning.

Table 1 shows the averages for each score. The average typicality and monotonicity scores (2.88 and 3.76, respectively) are each significantly greater than the chance threshold

of 2.0 ($t(40) = 5.90$, $t(40) = 19.31$, respectively, $p < .001$). (All tests are 2-tailed.) The average premise-probability score (2.71) also is reliably greater than 2 ($t(40) = 3.68$, $p < .01$), as is the average *PPP* score of 2.41 ($t(40) = 2.13$, $p < .05$). The average diversity score (2.34) is only marginally significant ($t(40) = 1.86$, $p = .07$).

There were 13, 12, and 16 3-year-olds, 4-year-olds, and 5-year-olds, respectively. A one-way ANOVA on age revealed no differences in any of the five scores we tabulated.

——- Insert Table 1 about here ——-

Discussion. The results for typicality and monotonicity suggest that the children grasped the character of the questions, and have accessible intuitions about argument strength. Conformity to *PPP* by the children in our sample is greater than chance, although not robust. With children between 3 and 5 years of age, many violations of *PPP* could represent inattention or inappropriate biases. Consequently, the noteworthy result of the study is not the number of violations of *PPP* by these young children, but rather the fact that conformity to it was statistically reliable. This is all the more impressive in view of the fact that credit for *PPP* required a choice of one premise pair in the PPP_{prob} item, and the opposite premise pair in the corresponding PPP_{stren} item.

In contrast, the children did not exhibit reliable conformity to the diversity principle inasmuch as their average diversity score was only marginally greater than expected by chance. According to the normative analysis presented earlier, a low diversity score does not signify immature inductive competence. It may signify nothing more than disagreement with adults about the relative probability of premise-sets. Such disagreement hinges on mere fact, not inductive logic. Thus, use of the diversity principle as a normative standard would have underestimated children's scientific competence in the present experiment. (Note, however, that the diversity scores were not reliably lower than the *PPP* scores, according to a correlated t-test.)

3.2 Experiment 2

As mentioned above, conformity to *PPP* in Experiment 1 required the child to make contrasting choices across corresponding PPP_{prob} and PPP_{stren} items. Response perseveration, therefore, cannot explain correct answers. It seemed prudent to ensure that the better-than-chance *PPP* scores in Experiment 1 were not due to the opposite bias, namely, to change

responses between corresponding PPP_{prob} and PPP_{stren} items. For this purpose, we replicated Experiment 1 with just one change. Conformity to *PPP* now required the child to choose the same clues across PPP_{prob} and PPP_{stren} items.

Participants. Preschool children from the same day care centers as in Experiment 1 served as subjects. In fact, children were randomly assigned to Experiment 1 versus 2. The sample included 19 boys and 19 girls (age range = 3.08 to 5.75 years; mean age = 4.31; SD = .73). One child was excluded from the analysis due to absence from the second interview session.

Materials, design, and procedure. The materials, design and procedure were identical to Experiment 1, with one exception. The instructions for PPP_{prob} items now read as follows.

> **Experiment 2 instructions for a PPP_{prob} item:** Detectives Max and Morgan found some clues. After they saw their clues, one of them said "I didn't think it would be like that. I am very surprised." But we don't know who said it. Max found that lions and tigers have thick blood. Morgan found that lions and rhinos have thick blood. Which one of them was surprised? Do you think Max was surprised because he found that lions and tigers have thick blood, or Morgan was surprised because he found that lions and rhinos have thick blood? (The order of presentation of the two sets of clues was individually randomized.)

Results. For each child, we defined typicality, monotonicity, premise-probability, diversity, and *PPP* scores, just as before. Table 2 shows the average scores. The average typicality, monotonicity, and premise-probability scores were similar to before. Of greater relevance is the average *PPP* score of 2.89, which is reliably better than the chance score of 2.0 ($t(36) = 5.62$, $p < .001$), but not reliably better than the *PPP* score in Experiment 1 ($F(1) = 3.53$, $p = .06$ by a one-way analysis of variance). The average diversity score (2.62) was also reliably better than chance ($t(36) = 3.65$, $p < .01$), although not reliably better than the average diversity score in Experiment 1 ($F(1) = 1.24$, $p = .27$ by one-way ANOVA). Observe that *PPP* scores are once again higher than diversity scores (although not significantly).

There were 13, 17, and 7 3-year-olds, 4-year-olds, and 5-year-olds, respectively. A one-way ANOVA on age revealed no differences in any of the five scores we tabulated.

——- Insert Table 2 about here ——-

Discussion. The present results show that the findings of Experiment 1 cannot be explained by a bias to switch responses between the PPP_{prob} and PPP_{stren} items. Any such tendency would have reduced the average *PPP* score whereas it was in fact greater in the present experiment than in Experiment 1. The methodology of Experiment 1 thus appears to offer the more conservative measure of conformity to *PPP*. It requires a response change between PPP_{prob} and PPP_{stren} items in order to be credited with success.

3.3 Experiment 3

To test the reliability and generality of the results of Experiment 1, we replicated it using a sample of Taiwanese children. Taiwan offers a linguistic contrast with the American sample that is relatively free of confounding differences in standard of living and industrial development.

Participants. Preschool and 3rd-to-5th grade children served as subjects. They were recruited from a day care center and an elementary school in a middle class neighborhood of Taichung, Taiwan. The preschool sample included 14 boys and 21 girls (age range = 3.75 to 5.83 years; mean age = 4.83; SD = .58). The elementary school sample included 41 boys and 27 girls (age range = 7.83 to 11.08; mean age = 9.71; SD = .92). One preschooler was excluded from the analysis due to absence from the second interview session.

Materials and Design. The design was identical to that of Experiment 1. All materials and instructions were translated into Chinese, and then verified via back-translation to English by a Taiwanese citizen unfamiliar with the purposes of the experiment. Two changes were then made. First, some categories were changed to adapt to Taiwanese culture. Specifically, Robins, Blue jays, and Broccoli were replaced by Pigeons, Canaries, and Bok Choi respectively. Second, instructions for PPP_{prob} items were partially modified. A pilot study suggested that some Taiwanese preschool children were not familiar with expressions such as "I am not surprised." Therefore, we substituted "which is more likely to

happen" for not being surprised. As in Experiment 1, augmenting the premise-probability score required choice of the homogeneous pair whereas augmenting the diversity score required choice of the heterogeneous pair. Credit for *PPP* thus presupposed a change in response.

Procedure. The procedure was identical to that used in Experiment 1. All children were interviewed by a native Taiwanese experimenter.

Results and discussion. For each child, we assigned typicality, monotonicity, premise-probability, diversity, and *PPP* scores as in Experiments 1 and 2. Averages are shown in Table 3. All five scores from each of the preschool and elementary school groups were significantly greater than the chance score of 2.0. The results with Taiwanese children thus suggest that conformity to *PPP* is not dependent on the use of English to frame and answer questions about argument strength.

——- Insert Table 3 about here ——-

A one-way ANOVA was performed to test for age effects on the five scores. Third-to-fifth-grade children had significantly higher typicality and premise-probability scores compared to preschool children ($p < .001$). No other differences were significant.

Comparing the preschool Taiwanese children with the American preschoolers in Experiment 1 reveals no reliable difference in their *PPP* scores. (The different methodology of Experiment 2 precludes its comparison with the Taiwanese sample.) The difference in diversity scores between Experiments 1 and 3 is only marginally significant ($F(1) = 3.64$, $p = .06$ by one-way ANOVA).

3.4 Experiment 4

The children who participated in Experiments 1 - 3 showed reliable conformity to *PPP*, but nonetheless violated it on numerous occasions. Pooling across experiments, there were 180 subjects, each of whom had four opportunities to make responses that respect the principle. Of these 720 opportunities, 67% were consistent with *PPP*. The contrary responses might represent loss of focus on the task, misremembered premises, or other lapses. To simplify the task, we decided to drop the detective story context on the grounds that it might be

distracting the children more than motivating them. We therefore replicated Experiment 1 with more straightforward queries about probability and argument strength. The replication also raised to six the number of times conformity to *PPP* could be assessed. As in Experiments 1 and 3, conformity required contrasting choices in corresponding PPP_{prob} and PPP_{stren} items.

Participants. Preschool children were recruited from two day care centers in a middle class neighborhood of Houston, Texas. (The day care centers were different from those in Experiments 1 - 2). The sample included 29 boys and 37 girls (age range = 3.58 to 5.67 years; mean age = 4.59; SD = .56) in the experiment. One child was excluded from the analysis due to absence from the second interview session.

Design. A total of 20 questions were presented to each child. They were (a) four typicality items, (b) four monotonicity items, (c) six PPP_{prob} items, and (d) six PPP_{stren} items. The six PPP_{stren} items involved the same premises as the six PPP_{prob} items, just as for Experiments 1 - 3. Also like the previous experiments, questions were randomly distributed over two testing sessions (on different days) with the constraint that neither of the sessions for any child include both a PPP_{prob} and PPP_{stren} question related to the same argument. See Appendix B for the stimuli used.

Materials and procedure. The materials and procedure were identical to those of Experiments 1 to 3 except that the instructions for the PPP_{prob} and PPP_{stren} items were shortened to the following.

> **Experiment 4 instructions for a PPP_{prob} item:** Suppose you are curious about animals, so you decide to study them. What do you think is more likely to happen? Is there a better chance that lions and tigers have thick blood, or that lions and rhinos have thick blood? (The order of presentation of the two sets of clues was individually randomized.)
>
> **Experiment 4 instructions for a PPP_{stren} item:** Suppose you are curious about animals, so you decide to study them. If you want to know whether all animals have thick blood, which would be the better clue, that lions and tigers

have thick blood, or that lions and rhinos have thick blood? (The order of presentation of the two sets of clues was individually randomized.)

Results and discussion. Typicality and monotonicity scores range from 0 to 4, and premise-probability, diversity, and *PPP* scores from 0 to 6. Table 4 shows average scores. Children performed significantly better than chance on all scores except diversity. Indeed, diversity scores were reliably *worse* than chance performance ($t(64) = -2.54$, $p < .05$). Use of the diversity principle instead of *PPP* would thus have underestimated inductive competence in the present experiment.

——- Insert Table 4 about here ——-

In Experiment 4 there were 9, 35, and 21 3-year-olds, 4-year-olds, and 5-year-olds, respectively. Surprisingly, younger subjects tended to have higher scores than older ones. Three-year-olds had the highest *PPP* scores (average 3.89) followed by 4-year-olds (average 3.57) and 5-year-olds (average 2.86). These differences are marginally significant by a one-way ANOVA on age ($F(2) = 2.20$, $p = .12$). The diversity scores for 3-, 4-, and 5-year-olds were 3.00, 2.86, and 1.90, respectively. A one-way ANOVA for age on this variable is significant ($F(2) = 4.05$, $p < .05$). The typicality scores for 3-, 4-, and 5-year-olds were 3.11, 2.71, and 2.38, respectively. A one-way ANOVA reveals this difference to be marginally significant ($F(2) = 2.99$, $p = .06$). No other age differences in Experiment 4 approached signficance.

Only Experiment 1 is comparable to Experiment 4 since both involved American subjects and the same requirements for conformity to *PPP*. Unlike the contrast between Experiments 1 and 2, however, subjects were not assigned randomly to Experiments 1 versus 4. Statistical inferences concerning the latter experiments are therefore somewhat tenuous. To compare the results of Experiments 1 and 4 as best as possible, we first converted scores to percentages. On this basis, conformity to *PPP* was lower in the present experiment than before. The average percentage for *PPP* was .60 in Experiment 1 and .56 in Experiment 4. One-way analysis of variance, however, revealed no significant difference between the two experiments in average typicality, monotonicity, premise-probability, and *PPP* scores (converted to percents). In contrast, the average diversity score was reliably lower in the present experiment ($F(1) = 9.52$, $p < .01$).

There was no significant interaction in a two-way ANOVA involving age (3, 4, or 5 years) and experiment (1 versus 4). ($F(2) = 2.02$, $p = .14$.) The difference between the two experiments in overall performance on *PPP* seems nonetheless to derive from the 5-year-olds. Indeed, the average (percentage) *PPP* score in Experiment 1 for the 5-year-olds was .65 but only .48 in Experiment 4. This difference between the 5-year-old's performance in the two experiments is marginally signficant by a one-way ANOVA ($F(1) = 3.50$, $p = .07$.) The 5-year-olds in the two experiments differed even more in diversity score. In Experiment 1, they scored 67% on the average, whereas this drops to 32% in Experiment 4. By a one-way ANOVA, this difference is reliable ($F(1) = 18.30$, $p < .001$.) Finally, there was a marginally significant decrease in the 5-year-old's monotonicity score, from 97% in Experiment 1 to 87% in Experiment 4. ($F(1) = 3.89$, $p = .06$.) The lower monotonicity score suggests that some of the 5-year-olds in Experiment 4 did not understand the task (possibly because the detective-story frame was dropped). For the younger subjects (3 and 4 years), there was no reliable difference in any of the scores between Experiments 1 and 4. In particular, the average *PPP* score for 3- and 4-year-olds in Experiment 1 is .57 whereas it is .61 in Experiment 4.

Overall, our attempt to simplify the questions by suppressing the detective story was not successful. Respect for *PPP* was not enhanced under the present instructions, compared to the more elaborate version of Experiment 1. Indeed, both monotonicity and *PPP* scores tended to be lower for 5-year-olds in Experiment 4 than in Experiment 1. Although no single finding is conclusive in this connection, the present results suggest that inductive competence in American children is better assessed in a story context involving "clues" and related concepts.

Although *PPP* performance was weak, the children in Experiment 4 nonetheless manifested statistically reliable conformity to *PPP*, like the children in Experiments 1 - 3. The performance of the 3- and 4-year-olds in Experiment 4 was especially striking in this regard.

3.5 Experiment 5

Although children in the foregoing studies showed conformity to *PPP*, their performance was less than perfect. To document the developmental change that leads to more robust appreciation of *PPP*, we posed similar questions to a sample of adults. We included a range of questions about argument strength, not all of which were relevant to the present study

(e.g., concerning the evaluation of arguments with specific conclusions). In what follows, we limit attention to typicality, monotonicity, and tests of *PPP*.

Participants. A group of 45 American college students from Rice University participated in the experiment. They ranged in age from 18.50 to 23.40 years (mean age = 20.00, SD = 1.18). There were 27 females and 18 males.

Materials and Design. A pretest of category familiarity was followed by typicality, monotonicity, PPP_{prob} and PPP_{stren} items. The stimuli figuring in the study are shown in Appendix C.

Procedure. Compared to the experiments with children, the testing procedure in the present experiment was modified in the following ways. (a) All questions were posed in the same session, rather than across two sessions. (b) No pictures were shown. (c) Instead of embedding questions within the detective story, subjects were asked "which is more likely?" for the PPP_{prob} items and "which gives you better reason to believe?" for the PPP_{stren} items. (d) Subjects received 2 typicality items and 2 monotonicity items, along with 3 PPP_{prob} items matched with 3 corresponding PPP_{stren} items. Finally, (e) subjects were tested in a group setting using booklets; questions were presented in one of three random orders.

Results and discussion. Subjects scored significantly better than chance ($p < .001$) on all measures (see Table 5). Seventy-six percent of the subjects were consistent with *PPP* on all three occasions. The results suggest that adult inductive intuition conforms strongly to *PPP*, at least in a typical American university.

——- Insert Table 5 about here ——-

4 General Discussion

We have argued that the *Premise Probability Principle* (or *PPP*) has greater normative justification than the *diversity principle* evoked in previous studies. The *PPP* is therefore the better criterion of mature inductive methodology. In Experiment 1, American preschoolers showed reliable (albeit imperfect) conformity to *PPP*. Indeed, they showed

more conformity to *PPP* than to the diversity principle. Experiment 2 revealed that conformity to *PPP* is not diminished by requiring the same argument to be selected in both the PPP_{prob} and PPP_{stren} items. In fact, mature responding increased under this requirement. In Experiment 3, it was seen that Taiwanese children showed as much conformity to *PPP* as their American counterparts. Experiment 4 explored the consequences for American children of removing the detective-game context used in Experiments 1 - 3. Conformity to *PPP* was still detectable under these conditions, but weaker than before, especially among 5-year-olds. Finally, Experiment 5 revealed consistent application of *PPP* among a sample of American college students. We now discuss the bearing of these findings on the question of whether children can be properly conceived as "little scientists."

Assimilating cognitive development to rational scientific inquiry is a bold and interesting enterprise, recently articulated in Gopnik and Meltzoff (1996).[9] Conceiving the child as scientist, however, requires more than recording the impressive theories that he or she constructs along the way to adulthood. Intellectual growth must also be shown to derive from the kind of theory construction attributed to sensible scientists. Otherwise, the succession of theories that enter the child's mind can be equally well attributed to triggering (involving no evidential justification) in the well-known sense described by Chomsky (1986, 1988) for linguistic development. To illustrate, whether children are little scientists is a separate issue from whether they have a semblance of mature biological theory.[10] Whatever theory young children reach at a given age, there remains the methodological question of how they become convinced of its truth.

Giving content to the thesis of child-as-scientist therefore requires substantive assumptions about rational inductive methodology. We have seen that the diversity principle does not serve this purpose since its normative status is open to doubt. The principle of premise probability (*PPP*) has better credentials since it follows from the axioms of probability and a plausible definition of argument strength. Our experiments reveal tenuous but detectable conformity to *PPP* on the part of young children. If these findings are supported and extended by further experimentation, they provide one clear sense in which young children's inductive methodology can be interpreted as properly scientific.

To formulate a more general theory of young children's scientific methodology, it is tempting to credit them with all sufficiently simple principles that follow from the Bayesian analysis of empirical inquiry. One difficulty with such a proposal is that Bayesians do not

speak with a single voice. Differences are especially pronounced when it comes to analyzing how beliefs should change under the impact of evidence.[11] Bayesianism appears to be on firmer ground when its claims are limited to the probabilities assigned by a rational agent at a given moment of time. Such "synchronic" Bayesianism implies *PPP* and many other compelling principles. An attractive theory of scientific competence may thus be stated as follows.

> (*) *Children as Synchronic Bayesians.* Young children's judgments of probability and evidence conform to every (sufficiently simple) consequence of standard accounts of synchronic Bayesianism.

Among the "sufficiently simple" consequences of synchronic Bayesianism is *PPP*, as we have seen. Other simple consequences include:

> Comparative strength principle (*CSP*): Let arguments $A, B \;/\; D$ and $A, C \;/\; D$ be given, where (1) D logically implies A, B, and C, and (2) B and C have equal probability. Then the strength of the first argument is greater than the strength of the second if and only if the strength of the argument $A \;/\; C$ is greater than the strength of $A \;/\; B$.

The following arguments illustrate the principle.

(k)
Lions have leukocytes inside. (A)
Horses have leukocytes inside. (B)
———————————
All animals have leukocytes inside. (D)

(l)
Lions have leukocytes inside. (A)
Tigers have leukocytes inside. (C)
———————————
All animals have leukocytes inside. (D)

(m)
Lions have leukocytes inside. (A)
———————————
Tigers have leukocytes inside. (C)

(n)
Lions have leukocytes inside. (A)
———————————
Horses have leukocytes inside. (B)

Because the predicate is blank, B and C have equal probability. *CSP* thus asserts that (k) is stronger than (l) if and only if (m) is stronger than (n). The descriptive theory based

on synchronic Bayesianism thus implies that young children's judgments of strength will conform to this same biconditional. We find this descriptive claim plausible.

Unfortunately, other consequences of (*) are likely to be false. One simple implication of synchronic Bayesianism is the *conjunction principle*, according to which $\mathrm{P}(p) \geq \mathrm{P}(p \wedge q)$ for any statements p and q. There is considerable evidence that adult reasoners often violate the conjunction principle.[12] Several developmental studies point to the same pattern of fallacious reasoning in children, both American and Taiwanese (Davidson, 1995; Sides, Lo, Schweingruber, and Osherson, 2001). It thus seems necessary to weaken Theory (*). We suggest the following.

> (**) *Children as Synchronic Bayesians (revised).* For every consequence C of synchronic Bayesianism, if adults' judgments conform to C, then so do children's.

Theory (**) is supported by our results on *PPP*, and appears not to be contradicted by other available data. If it survives further test, the theory would provide only partial vindication of the "little scientist" thesis. For one thing, (**) bears on just a fraction of the full complement of principles involved in sound scientific practice. Moreover, adult reasoners seem to be prey to such a diverse array of fallacies that it may be overly generous to qualify most people of any size as genuine scientists.[13] To the extent that (**) is true, it only supports equating the methodological prowess of children and adults, allowing for defects at all ages. Instead of conceiving of children as diminuitive yet astute methodologists, theory (**) would thus sustain a perspective that might be termed the "equal scientist" thesis of cognitive development.

References

Ahn, W., Kalish, C. W., Medin, D. L., and Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54:299–352.

Atran, S. (1995). Classifying nature across cultures. In Smith, E. E. and Osherson, D., editors, *Invitation to Cognitive Science (2nd Edition): Thinking*, pages 131 – 174. Cambridge MA, MIT Press.

Bar-Hillel, M. (1991). Commentary on Wolford, Taylor, and Beck: The conjunction fallacy? *Memory and Cognition*, 19(4):412 – 417.

Bjorklund, D. F., Thompson, B. E., and Ornstein, P. A. (1983). Developmental trends in children's typicality judgments. *Behavior Research Methods & Instruction*, 15(3):350–356.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge MA, MIT Press.

Choi, I., Nisbett, R. E., and Smith, E. E. (1997). Culture, Categorization and Inductive Reasoning. *Cognition*, 65:15 – 32.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use*. New York NY, Praeger.

Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*. Cambridge MA, MIT Press.

Coley, J. D. (1995). Emerging Differentiation of Folkbiology and Folkpsychology: Attributions of Biological and Psychological Properties to Living Things. *Child Development*, 66:1865 – 1874.

Davidson, D. (1995). The Representativeness Heuristic and the Conjunction Fallacy Effect in Children's Decision Making. *Merrill-Palmer Quarterly*, 41(3):328 – 346.

Dawes, R., Mirels, H. L., Gold, E., and Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science*, 4(6):396–400.

Dulany, D. E. and Hilton, D. J. (1991). Conversational implicature, conscious representation, and the conjunction fallacy. *Social Cognition*, 9:85 – 110.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge MA, MIT Press.

Eells, E. (1985). Problems of Old Evidence. *Pacific Philosophical Quarterly*, 66:283–302.

Fitelson, B. (1999). The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity. *Philosophy of Science*, 66:S362–S378.

Fraassen, B. V. (1988). The Problem of Old Evidence. In Autin, D., editor, *Philosophical Analysis*, pages 153 – 65. Dordrecht, Kluwer Academic Publishers.

Franklin, A. and Howson, C. (1984). Why Do Scientists Prefer to Vary Their Experiments? *Studies in the History and Philosophy of Science*, 15:51–62.

Garber, D. (1983). Old Evidence and Logical Omniscience in Bayesian Confirmation Theory. In Earman, J., editor, *Testing Scientific Theories*, pages 99 – 131. Minneapolis MN, University of Minnesota Press.

Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond 'heuristics and biases'. In Stroebe, W. and Hewstone, M., editors, *European Review of Social Psychology (Vol. 2)*, pages 83–115. Chichester, England, Wiley.

Gigerenzer, G. (1996). Reply to Tversky and Kahneman. *Psychological Review*, 103(3):592–3.

Glymour, C. (1980). *Theory and Evidence*. Princeton NJ, Princeton University Press.

Goldstone, R. L. and Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, 65:231–262.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiments. *Journal of the Royal Statistical Society*, B 22:319–331.

Gopnik, A. and Meltzoff, A. (1996). *Words, Thoughts, and Theories*. Cambridge MA, MIT Press.

Gustason, W. (1994). *Reasoning from Evidence: Inductive Logic*. New York City, Macmillan.

Gutheil, G. and Gelman, S. (1997). Children's Use of Sample Size and Diversity Information within Basic-Level Categories. *Journal of Experimental Child Psychology*, 64:159–174.

Gutheil, G., Vera, A., and Keil, F. (1998). Do houseflies think? Patterns of induction and biological beliefs in development. *Cognition*, 66:33–49.

Hatano, G. and Inagaki, K. (1994). Young children's naive theory of biology. *Cognition*, 50:171–188.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In Oaksford, M. and Chater, N., editors, *Rational models of cognition*, pages 248–274. New York NY, Oxford University Press.

Heit, E. and Hahn, U. (1999). Diversity-Based Reasoning in Children Age 5 to 8. In *Proceedings of the Twenty-First Annual Conference of the Cognitive Science Society*, Mahawah NJ. Erlbaum.

Heit, E. and Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:411–422.

Hempel, C. G. (1966). *Philosophy of Natural Science*. Englewood Cliffs NJ, Prentice Hall.

Hertwig, R. and Chase, V. M. (1998). Many Reasons or Just One: How Response Mode Affects Reasoning in the Conjunction Problem. *Thinking and Reasoning*, 4(4):319–352.

Hertwig, R. and Gigerenzer, G. (1999). The 'Conjunction Fallacy' Revisited: How Intelligent Inferences Look Like Reasoning Errors. *Journal of Behavioral Decision Making*, 12:275–305.

Hintikka, J. (1968). The varieties of information and scientific explanation. In van Rootselaar, B. and Staal, J. F., editors, *Logic, Methodology and Philosophy of Science III*, pages 311–331. Amsterdam, North-Holland.

Horwich, P. (1982). *Probability and Evidence*. Cambridge, Cambridge University Press.

Howson, C. and Urbach, P. (1993). *Scientific Reasoning: The Bayesian Approach*. La Salle, IL, Open Court.

Jeffrey, R. (1992). *Probability and the Art of Judgment*. Cambridge UK, Cambridge University Press.

Johnson, S. and Solomon, G. (1997). Why Dogs Have Puppies and Cats Have Kittens: The Role of Birth in Young Children's Understanding of Biological Origins. *Child Development*, 68(3):404–419.

Kahneman, D. and Tversky, A. (1996). On the Reality of Cognitive Illusions. *Psychological Review*, 103(3):582–591.

Kaplan, M. (1996). *Decision Theory as Philosophy*. Cambridge UK, Cambridge University Press.

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

Keil, F., Smith, W. C., Simons, D. J., and Levin, D. T. (1998). Two dogmas of conceptual empiricism: implications for hybrid models of the structure of knowledge. *Cognition*, 65:103 – 135.

Koslowski, B. (1996). *Theory and Evidence: The development of Scientific Reasoning*. Cambridge MA, MIT Press.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, 96(4):674–689.

Levi, I. (1991). *The Fixation of Belief and Its Undoing*. Cambridge UK, Cambridge University Press.

López, A. (1995). The diversity principle in the testing of arguments. *Memory & Cognition*, 23(3):374 – 382.

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

López, A., Gelman, S. A., Gutheil, G., and Smith, E. E. (1992). The Development of Category-based Induction. *Child Development*, 63:1070–1090.

Mahler, P. (1993). *Betting on Theories*. Cambridge UK, Cambridge University Press.

Markman, E. (1989). *Categorization and naming in children: Problems of induction*. Cambridge MA, MIT Press.

Martin, E. and Osherson, D. (Spring, 1998). *Elements of Scientific Inquiry*. Cambridge MA, MIT Press.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago IL, University of Chicago press.

Myrvold, W. C. (1996). Bayesianism and Diverse Evidence: A Reply to Andrew Wayne. *Philosophy of Science*, 63:661 – 665.

Osherson, D. (1995). Probability judgment. In Osherson, D. and Smith, E. E., editors, *Invitation to Cognitive Science: Thinking (Second Edition)*. Cambridge MA, M.I.T. Press.

Osherson, D., Smith, E., Wilkie, O., López, A., and Shafir, E. (1990). Category-Based Induction. *Psychological Review*, 97(2):185–200.

Osherson, D., Smith, E. E., and Shafir, E. (1986). Some origins of belief. *Cognition*, 24:197–224.

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

Rips, L., Shoben, E., and Smith, E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 12:1 – 20.

Shafir, E., Smith, E., and Osherson, D. (1990). Typicality and reasoning fallacies. *Memory and Cognition*, 18(3):229–239.

Sides, A., Lo, Y., Schweingruber, H., and Osherson, D. (2001). The development of probabilistic reasoning. To appear.

Sides, A., Osherson, D., Bonini, N., and Viale, R. (in press). On the reality of the conjunction fallacy. *Memory & Cognition*.

Skyrms, B. (1986). *Choice and Chance: An Introduction to Inductive Logic (3rd Edition)*. Belmont CA, Wadsworth.

Solomon, G., Johnson, S., Zaitchik, D., and Carey, S. (1996). Like Father, Like Son: Young Children's Understanding of How and Why Offspring Resemble Their Parents. *Child Development*, 67:151–171.

Springer, K. (1992). Children's beliefs about the biological implications of kinship. *Child Development*, 63:950–959.

Springer, K. (1996). Young Children's Understanding of a Biological Basis for Parent-Offspring Relations. *Child Development*, 67:2841–2856.

Springer, K. and Keil, F. (1989). On the development of biologically specific beliefs: The case of inheritance. *Child Development*, 60:637–648.

Springer, K. and Keil, F. (1991). Early differentiation of causal mechanisms appropriate to biological and nonbiological kinds. *Child Development*, 62:767 – 781.

Tversky, A. and Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90:293–315.

Viale, R. and Osherson, D. (2000). The diversity principle and the little scientist hypothesis. *Foundations of Science*, 5(2):239–253.

Wayne, A. (1995). Bayesianism and Diverse Evidence. *Philosophy of Science*, 62:111 – 121.

Wolford, G., Taylor, H. A., and Beck, J. R. (1990). The conjunction fallacy? *Memory and Cognition*, 18(1):47 – 53.

Notes

(1) Arguments (g) and (h) are inspired by an example in López, Atran, Coley, Medin and Smith (1997). Observe that the predicate in the present example is just as blank as *has an ulnar artery*, used in other studies (e.g., López, 1995). Indeed, many animals (like sponges and jellyfish) will be supposed not to possess such an artery, whereas any of them could potentially be host to the Floxum parasite. Application of the Floxum predicate thus appears to be even less determinate than that of the ulnar predicate (hence the former is even blanker than the latter).

(2) This standard objection to evoking similarity in theories of reasoning is discussed in Osherson, Smith and Shafir (1986, p. 220), Shafir, Smith and Osherson (1990, p. 237), and Keil, Smith, Simons and Levin (1998, p. 107), among other places. A more favorable view of similarity in the theory of reasoning is offered by Goldstone and Barsalou (1998). The sensitivity of similarity based reasoning to the predicate in play is documented in Heit and Rubinstein (1994). In contrast to similarity based inference, results in Ahn, Kalish, Medin and Gelman (1995) underline the importance of explanatory theories to adult reasoning.

(3) Another example appears in Wayne (1995, pp. 115-6), who notes that before Maxwell electromagnetic and optical phenomena were considered diverse whereas today they are judged to be similar.

(4) Instead of trying to draw a sharp line between blank and meaningful predicates, increasing blankness might be postulated to increase the importance of diversity. According to this idea, the more blank the predicate, the more potent is diversity within arguments involving that predicate. But the idea seems not to be sustainable. Consider:

(o)
Minneapolis had a drop in crime last year.
St. Paul had a drop in crime last year.

All major U.S. cities had a drop in crime last year.

(p)
Minneapolis had a drop in crime last year.
San Diego had a drop in crime last year.

All major U.S. cities had a drop in crime last year.

The predicate *had a drop in crime last year* is highly meaningful. Yet Minneapolis and San Diego seem more diverse than Minneapolis and Saint Paul, in correlation with the greater strength of (p) compared to (o). This is to be contrasted with the lack of correlation between diversity and strength in arguments (g), (h) above, involving the highly blank predicate *often carry the parasite Floxum*. Hence, greater blankness in predicates does not coincide in general with a stronger relation between strength and diversity. In any event, we shall shortly argue that diversity, *per se*, is not the relevant variable governing strength even when these two attributes of arguments *are* correlated.

(5) Recall that $\mathrm{P}(C \mid A \wedge B)$ denotes the conditional probability of C given A and B. The symbol "$\wedge$" denotes conjunction. The unconditional probability of the conclusion C (namely, without assuming A and B), is denoted by $\mathrm{P}(C)$. Many authors have measured argument strength by comparing the prior probability of the conclusion with its probability given the premises, e.g., Good (1960), Hintikka (1968).

(6) See Fitelson (1999) for measures of argument strength that are alternative to the ratio definition relied on here. (Most definitions leave *PPP* intact.) In the current discussion, premise diversity is taken to be a primitive concept, not defined in terms of probability and confirmation. In contrast, Myrvold (1996) and Wayne (1995) (among others) attempt to reconstruct the concept of diversity in terms of probability and confirmation. Within such an analysis, "diversity" (as reconstructed) is constrained to be systematically related to confirmation, but this fact does not bear on the diversity principle as interpreted here. See Heit (1998) and Viale and Osherson (2000) for further discussion.

(7) In their pioneering study, López et al. (1992) use the following method to assess the relative strength of two arguments. Given animal categories like *cat, buffalo*, and *cow*, children are told that cats and buffalos possess one blank property (e.g., having leukocytes inside) whereas cats and cows possess another blank property (e.g., having ulnaries inside). They are then queried: "Do you think *all animals* have what cats and buffalos have inside or what cats and cows have inside?" The choice of premise-set is taken to correspond to the stronger argument. This method is ingenious but we have not adopted it because the presence of distinct predicates for the two sets of premises risks influencing judgments of argument strength. Although the predicates are designed to be blank, their resemblance to words the child might know could provoke extraneous reasoning.

(8) There seems to be no normative justification for such a choice even though it is widespread among college students. The choice follows from *PPP* only if the more typical category is less likely to possess the predicate than the less typical category. When predicates are blank (as in most demonstrations of the typicality effect), it is hard to see what motivates the latter judgment. Most of the typicality contrasts appearing in our stimuli are documented for children and adults in Bjorklund, Thompson, and Ornstein (1983), and Rips, Shoben, and Smith (1973), The remaining contrasts were verified through pilot testing with college students.

(9) They write: "The cognitive processes that underlie science are similar to, or indeed even identical with, the cognitive processes that underlie much of cognitive development." (p. 32)

(10) For conflicting evidence and views about young children's biological theories, see Coley (1995), Gutheil, Vera and Keil (1998), Hatano and Inagaki (1994), Johnson and Solomon (1997), Solomon, Johnson, Zaitchik and Carey (1996), Springer (1992, 1996), and Springer and Keil (1989, 1991).

(11) For rival theories within the Bayesian family, see Jeffrey (1992), Horwich (1982), Levi (1991), Mahler (1993), and Kaplan (1996). Points of disagreement include the problem of "old evidence." Divergent opinions about this issue are available in van Fraassen (1988), Eells (1985), Garber (1983), Glymour (1980), Howson and Urbach (1993), and Kaplan (1996).

(12) Conjunction fallacies were first reported in Tversky and Kahneman (1983) but their reality has been contested. For elements of the debate, see Hertwig and Chase (1998), Dulany and Hilton (1991), Wolford, Taylor and Beck (1990), Bar-Hillel (1991), Hertwig and Gigerenzer (1999), Kahneman and Tversky (1996) and Gigerenzer (1996). Our view is that the fallacy is genuine since it can be demonstrated with two-choice alternatives under betting instructions involving real stakes. See Sides, Osherson, Bonini and Viale (2001).

(13) For example, naive reasoning appears to have difficulty interpreting conditional probability (Dawes, Mirels, Gold & Donahue, 1993). The conjunction fallacy and many similar phenomena could be cited in this connection.

Appendix A: Arguments in Experiments 1 - 3

The table below shows the arguments of Experiments 1 - 3. The same blank predicate was employed for arguments on a given line of the table. For each line, subjects chose which of the two arguments was stronger, or (in the case of PPP_{prob} items) which pair of premises is more likely to be true.

Predicates included: *have T cells, calcium, vitamin C, vitamin B-6, B-cells, thick blood, copper, iron, lutein, zinc, arteries*, and *antenna in/on their bodies*. The word "bug" was used when repeating statements involving the word "insect." For a random half of the subjects, the categories and predicates used in the monotonicity items were substituted for those of the PPP_{prob} and PPP_{stren} items, and vice versa. By "stronger argument," is meant the argument deemed stronger by most adults. Note that in Experiment 3 (involving Taiwanese children), pigeons were substituted for robins, canaries were substituted for blue jays, and bok choi was substituted for broccoli.

Item type	*Weaker argument*	*Stronger argument*
Typicality	Penguins / All birds	Robins / All birds
	Pineapples / All fruits	Peaches / All fruits
	Potatoes / All vegetables	Broccoli / All vegetables
	Sea horses / All sea animals	Sharks / All sea animals
Monotonicity	Rabbits / All animals	Rabbits & cows / All animals
	Dogs / All animals	Dogs & elephants / All animals
	Sheep / All animals	Sheep & goats / All animals
	Sparrows / All birds	Sparrows & ostriches / All animals
PPP_{prob}	Lions & tigers / All animals	Lions & rhinoceroses / All animals
and	Horses & zebras / All animals	Horses & mice / All animals
PPP_{stren}	Dolphins & whales / All sea animals	Dolphins & star fish / All sea animals
	Ladybugs & beetles / All insects	Ladybugs & ants / All insects

Appendix B: Arguments in Experiment 4

The table below shows the arguments of Experiment 4. Predicates included: *have T cells*, *calcium*, *vitamin C*, *vitamin B-6*, *B-cells*, *thick blood*, *copper*, *iron*, *lutein*, *zinc*, *arteries*, *protein*, *U cells*, and *vitamin K in their bodies* .

Item type	*Weaker argument*	*Stronger argument*
Typicality	Penguins / All birds	Robins / All birds
	Pineapples / All fruits	Peaches / All fruits
	Potatoes / All vegetables	Broccoli / All vegetables
	Sea horses / All sea animals	Sharks / All sea animals
Monotonicity	Rabbits / All animals	Rabbits & cows / All animals
	Horses / All animals	Horses & mice / All animals
	Pigs / All animals	Pigs & bats / All animals
	Foxes / All animals	Foxes & Gorillas / All animals
PPP_{prob}	Lions & tigers / All animals	Lions & rhinoceroses / All animals
and	Sheep & goats / All animals	Sheep & monkeys / All animals
PPP_{stren}	Dolphins & whales / All sea animals	Dolphins & star fish / All sea animals
	Dogs & wolves / All animals	Dogs & elephants / All animals
	Sparrows & blue jays / All birds	Sparrows & ostriches / All birds
	Ladybugs & beetles / All insects	Ladybugs & ants / All insects

Appendix C: Arguments in Experiment 5

The table below shows the arguments of Experiment 4. The same blank predicate was employed for every argument, namely, *has a nutritional requirement for X*.

Item type	*Weaker argument*	*Stronger argument*
Typicality	Penguins / All birds	Robins / All birds
	Skunks / All mammals	Bears / All mammals
Monotonicity	Snakes / All reptiles	Snakes & alligators / All reptiles
	Bass & flounder / All fish	Bass, flounder, & sharks / All fish
PPP_{prob}	Lions & tigers / All mammals	Lions & rhinoceroses / All mammals
and	Rabbits & squirrels / All mammals	Rabbits & cows / All mammals
PPP_{stren}	Apes & chimpanzees/ All mammals	Apes & giraffes / All mammals

Author notes

We thank Susan Gelman, Douglas Medin, Laura Macchi, Katya Tentori, and Riccardo Viale for comments on earlier versions of the paper. The helpful remarks of three reviewers for this journal are also gratefully acknowledged. We thank the following people for allowing us to work with the children under their supervision: Mindy Butler, Taffy Dawson, Heng-Mu Li, Mei-Yu Liu, April Kliebert, Zane Tigett, and Janice Thompson. In addition, Mindy Butler and Hua Feng offered excellent advice about stimuli and procedures. Correspondence to D. Osherson, Rice University, P.O. Box 1892, Houston TX 77251-1892. E-mail: osherson@rice.edu.

TABLE 1
Mean Scores and t-test Results ($N = 41$) from Experiment 1

Score-type	Mean (SD)	t-value	p
Typicality	2.88 (.95)	5.90	$< .001$
Monotonicity	3.76 (.58)	19.31	$< .001$
Premise-probability	2.71 (1.23)	3.68	$< .01$
Diversity	2.34 (1.17)	1.86	ns
PPP	2.41 (1.24)	2.13	$< .05$

Note. Each score gives the mean number of times (out of 4) that the children answered the way adults usually do. T-tests compare each mean with 2.0, the score expected if the child responded uniformly randomly to each question.

TABLE 2

Mean Scores and t-test Results ($N = 37$) from Experiment 2

Score-type	Mean (SD)	t-value	p
Typicality	3.08 (.72)	9.11	$< .001$
Monotonicity	3.81 (.62)	17.87	$< .001$
Premise-probability	2.76 (1.14)	4.04	$< .001$
Diversity	2.62 (1.04)	3.65	$< .01$
PPP	2.89 (.97)	5.62	$< .001$

Note. Each score gives the mean number of times (out of 4) that the children answered the way adults usually do. T-tests compare each mean with 2.0, the score expected if the child responded uniformly randomly to each question.

TABLE 3

Mean Scores and t-test Results from Experiment 3

Score-type	Mean (SD)	t-value	p
	Preschool ($N = 34$)		
Typicality	3.03 (.90)	6.64	$< .001$
Monotonicity	3.82 (.58)	18.47	$< .001$
Premise-probability	2.79 (.98)	4.74	$< .001$
Diversity	2.85 (1.13)	4.40	$< .001$
PPP	2.65 (1.18)	3.20	$< .01$
	3rd to 5th grade ($N = 68$)		
Typicality	3.74 (.61)	23.32	$< .001$
Monotonicity	3.88 (.41)	38.21	$< .001$
Premise-probability	3.72 (.71)	20.01	$< .001$
Diversity	2.87 (1.27)	5.64	$< .001$
PPP	2.77 (1.32)	4.79	$< .001$

Note. Each score gives the mean number of times (out of 4) that the children answered the way adults usually do. T-tests compare each mean with 2.0, the score expected if the child responded uniformly randomly to each question.

TABLE 4
Mean Scores and t-test Results ($N = 65$) from Experiment 4

Score-type	Mean (SD)	t-value	p
Typicality	2.66 (.80)	6.70	$< .001$
Monotonicity	3.63 (.55)	24.04	$< .001$
Premise-probability	3.49 (1.35)	2.95	$< .01$
Diversity	2.57 (1.37)	-2.54	$< .05$
PPP	3.38 (1.49)	2.09	$< .05$

Note. Each score gives the mean number of times that children answered the way adults usually do. Scores could range from 0 to 4 for typicality and monotonicity, and from 0 to 6 for premise-probability, diversity, and *PPP*. T-tests compare each mean with 2.0 for typicality and monotonicity, and 3.0 for premise-probability, diversity, and *PPP*, the score expected if the child responded uniformly randomly to each question.

TABLE 5
Mean Scores and t-test Results ($N = 45$) from Experiment 5

Score-type	Mean (SD)	t-value	p
Typicality	1.62 (.68)	6.14	$< .001$
Monotonicity	1.76 (.53)	9.62	$< .001$
Premise-probability	2.84 (.37)	24.36	$< .001$
Diversity	2.53 (.86)	8.512	$< .001$
PPP	2.56 (.89)	7.97	$< .001$

Note. Each score gives the mean number of mature responses. Scores could range from from 0 to 2 for typicality and monotonicity, from 0 to 3 for premise-probability, diversity, and *PPP*. T-tests compare each mean with 1.0 for typicality and monotonicity. They compare means with 1.5 for premise-probability, diversity, and *PPP*. These comparison values are the scores expected if the students responded uniformly randomly to each question.