THINKING & REASONING, 2007, 13 (4), 394–425

Psychology Press
Taylor & Francis Group

Negative evidence and inductive generalisation

Charles W. Kalish and Christopher A. Lawson
University of Wisconsin-Madison, Madison, WI, USA

How do people use past experience to generalise to novel cases? This paper reports four experiments exploring the significance on one class of past experiences: encounters with negative or contrasting cases. In trying to decide whether all ravens are black, what is the effect of learning about a non-raven that is not black? Two experiments with preschool-aged, young school-aged, and adult participants revealed that providing a negative example in addition to a positive example supports generalisation. Two additional experiments went on to ask which kinds of negative examples offer the most support for generalisations. These studies contrasted similarity-based and category-based accounts of inductive generalisation. Results supported category-based predictions, but only for preschool-aged children. Overall, the younger children showed a greater reliance on negative evidence than did older children and adults. Most things we encounter in the world are negative evidence for our generalisations. Understanding the role of negative evidence is central for psychological theories of inductive generalisation.

INTRODUCTION

To be truly useful, knowledge gained in one context must be generalised to new cases. The central question for research on inductive inference is how people use their prior knowledge to make novel predictions. Imagine a child learns that her pet dog becomes sick after eating chocolate. The challenge is to extend that knowledge: What other animals get sick from chocolate? What else will make her dog sick? The generalisations people make will be affected by the evidence they have. One type of evidence is negative or

Correspondence should be addressed to Charles W. Kalish, University of Wisconsin-Madison, Educational Psychology, Room 1067 EdSciences, 1025 W. Johnson Street, Madison, WI 53706, USA. E-mail: cwkalish@wisc.edu

Thanks are due to Evan Heit and Vladimir Sloutsky for comments on a previous draft of this manuscript. This research was supported by NICHD grant R01 HD37520 to the first author.

http://www.psypress.com/tar DOI: 10.1080/13546780701273402

contrasting cases. In generalising about dogs and chocolate poisoning, what is the relevance of learning that chocolate is harmless to people?

The purpose of this study is to explore the role of negative evidence in children's and adults' inductive inferences. The quality of evidence can be defined relative to the conclusion or target in question. If the question is whether a dog (or dogs) will get sick given chocolate, positive evidence would be other category members who had the property (dogs with chocolate poisoning). Negative evidence would be non-members who lack the property (non-poisoned non-dogs). The most famous discussion of negative evidence is Hempel's (1945) raven's paradox. What supports the conclusion "All ravens are black"? Hempel notes that the conclusion is logically equivalent to "All non-black things are non-ravens". A white swan supports the latter conclusion in just the same way as a black raven supports the former. Since we should be committed to logical implications of our beliefs, we should be more likely to agree all ravens are black after seeing a white swan. Yet intuitively the negative evidence is not compelling. Indeed, one might imagine that discovering some birds are not black might make one *less* confident that all ravens are black. As Hempel's raven's paradox illustrates, the significance of negative evidence is a complex issue. It is not clear whether negative evidence supports or undermines a conclusion, it is not clear how much weight people give such evidence, nor is it clear how much weight is appropriate (normative).

The role of negative evidence in inductive generalisation is an important question in its own right; most of the things we encounter in the world are negative evidence for our hypotheses. Negative evidence is also significant as a test of competing accounts of inductive generalisation. Research on inductive inference has been largely limited to considering how people use positive evidence to make judgements. In the remainder of this introduction we identify predictions regarding negative evidence of three major accounts of inductive inference: Bayesian, category/relevance, and similarity models. Of particular interest are predictions about developmental changes in the processes of inductive inference and their implications for the significance of negative evidence.

Bayesian inference

The most compelling analysis of negative evidence and Hempel's paradox involves conditional probability and Bayesian inference (Howson & Urbach, 1993; McKenzie & Mikkelson, 2000; Nickerson, 1996). The basic insight is that negative evidence is much more likely than positive; there are many more non-black non-ravens than black ravens in the world. The chance that a given object is a negative instance is about the same whether or not the hypothesis under consideration (ravens are black) is true or false.

Thus the presence of negative cases is not very informative about the hypothesis. This analysis underlies Oaksford and Chater's (1994) Optimal Data Selection theory of hypothesis testing. People do not seek out counter-examples when testing hypotheses (e.g., in the Wason selection task) because most such examples are uninformative negative evidence. Looking for non-black non-ravens is a poor way to discover if all ravens are black. Similarly, syllogisms with negated premises are treated differently from those with positive premises (Evans & Handley, 1999). Oaksford, Chater, and Larkin (2000) argue that intuitions about the relative frequency of positive and negative instances can account for these effects as well.

If negative cases are seen as poor tests of hypotheses, then presumably their inclusion as evidence should have only a weak effect on conclusions. Consistent with this prediction, research on covariation detection reveals that negative evidence (cause absent/effect absent, or Cell D) has the weakest effect on judgements of association (Kao & Wasserman, 1993). The Bayesian prediction, then, is that provision of negative evidence should have a negligible impact on inductive generalisations (see McKenzie & Mikkelsen, 2005). It is important to note that the Bayesian perspective emphasises that the source of evidence is critical. The significance of a piece of evidence cannot be specified independent of how one came to learn that piece of evidence (Eells, 1982).

Although research on hypothesis testing suggests that adults consider the likelihood of evidence when making inductive judgements (McKenzie, Ferreira, Mikkelsen, McDermott, & Skrable 2001; Oaksford & Moussakowski, 2004), it is less clear whether young children will show similar effects. Young children appear insensitive to some basic statistical qualities of evidence. Preschool-aged children do not consider the amount of evidence available when making inductive judgements (Gutheil & Gelman, 1997; Lopez, Gelman, Gutheil, & Smith, 1992). Initial studies also indicated that young children did not appreciate the significance of diversity in evidence (Lopez, et al., 1992). They are just as likely to conclude that all birds have some novel property when given evidence about two very similar kinds of birds (e.g., robins and sparrows) as when given evidence about very diverse birds (e.g., robins and flamingos). Subsequent research has produced evidence for diversity, although the significance of these findings remains a matter of debate (see Gelman, 2004; Heit & Hahn, 2001). Other researchers (Lo, Sides, Rozelle, & Osherson, 2002) reinterpret diversity in terms of the probability of evidence. The diverse examples are less likely than the non-diverse if the conclusion is false; thus getting diverse examples provides stronger support for the conclusion. Preschool-aged children do indicate that people with low-probability evidence are in a better position to draw a conclusion about a general category than are people with high-probability evidence (Lo et al., 2002). This work has been restricted to judgements about

variations in positive evidence, instances with the property in question. It remains an open question whether or how the principle of probability of evidence will be applied to negative cases. The basic prediction, however, is that negative cases are equally likely whether or not the conclusion is true and should have little impact on inductive inferences.

Categorisation and relevance

In the literature on categorisation and word learning, the inductive problem is to identify the scope of a label or property. Given a black raven the challenge is to identify the category of black things (just this raven, a sub-set of ravens, all ravens, all birds?). Negative examples are foils that indicate the boundary of the category. A white swan contrasts at the species level, suggesting the category "raven". For identifying a category, negative examples may be just as significant and useful as positive.

Early research suggested that people were under-sensitive to negative instances (Johnson, 1972). It is somewhat more difficult to learn a concept given only negative examples than given only positive ones (Toppino & Johnson, 1974), but it is also easier to learn a concept given some negative information than given only positive information (Williams & Carmine, 1981). People use non-instances to constrain hypotheses about category boundaries and relevant attributes (Houtz, Moore, & Davis, 1973). More recent research has emphasised that concepts exist as parts of contrasting sets. Billman and Devilla (2001), for example, found that information about one category affects how a second, contrasting, category is learned. A similar perspective is evident in developmental studies. Children compare objects labelled with different words to discover distinctive features (Au & Laframboise, 1990; Waxman & Klibanoff, 2000); one label serves as negative evidence regarding the extension of the other. Young children are able to identify complementary classes from negative examples. However, the classes identified are not logical complements, but are pragmatic contrasts. For example, when asked to find dolls that are not fathers, children will select the mother dolls, but not the child dolls (Feldman, 1972). This suggests that negations are interpreted as low-frequency contrasting classes ("mothers"), not as high-frequency complement classes ("non-fathers").

To the extent that negative evidence has been addressed in the categorisation and induction literature the negations have been implicit (Evans & Handley, 1999): items are characterised as "white swans" rather than "non-black non-ravens". Oaksford et al. (2000) note that implicit negations can be seen as low-probability occurrences. Although non-black non-ravens are common, white swans are relatively rare. For inductive inference, though, the critical point is that negative cases are equally

frequent whether or not the hypothesis is true (the number of white swans is independent of the number of black ravens). However, absolute frequency may not be the only feature that can determine the likelihood of encountering a piece of evidence. Negative cases are particularly good pieces of evidence if one assumes a helpful teacher. By principles of relevance (Sperber & Wilson, 1995) one might expect that negative instances are exemplars falling just outside the category boundary. If all ravens are black, a teacher might present a white swan as a foil. A teacher is unlikely to use a white swan to illustrate the point that not all ravens are black. Negative evidence seems to be very significant for people's inductive inferences, at least about category membership and labelling. Perhaps this is because people interpret negative cases as "near-miss" foils that identify a relevant category.

Similarity-based inference

A final perspective on negative evidence relies on relative similarity judgements. This analysis has been applied to triad tasks involving negative evidence (Sloutsky & Fisher, 2004). For example, participants encounter a black raven and a white swan and are asked to predict the colour of a novel raven (see Gelman & Markman, 1986). Similarity-based accounts treat this kind of task as a forced-choice problem. The challenge is to decide which piece of evidence is more similar to the unknown target. The negative evidence is a competitor with the positive. The impact of the negative evidence depends on the difference between the similarity of the positive evidence to the target, and the similarity of the negative evidence to the target. This relation is formalised in several models of inductive inference, including Sloutsky and Fisher's (2004) SINC model.

Similarity-based inference may be characteristic of young children's inductive reasoning. As discussed above, preschool-aged children seem not to use evidence to form a category (Lopez et al., 1992) or evaluate a hypothesis. Rather, young children's property projections may be best understood as a process of similarity matching. Known cases are individually compared to the novel one (Sloutsky & Fisher. 2004). What determines whether a property is projected to a novel case is how similar that case is to ones known to possess the property (see Kahneman & Frederick, 2005). Dissimilarity to cases known to lack the property also supports inductive projection. Like the category-based accounts, the relative similarity models predict that negative evidence will have a significant impact on inductive judgement. The two models diverge, however, in their predictions about just which kinds of negative evidence will be most supportive of generalisations. These differential predictions are discussed with respect to Experiments 2 and 3 below.

Summary

The general question of induction is how people use information about known cases to make predictions about novel ones. Known cases may stand in various kinds of relations to the novel ones; in particular, some known cases may be positive evidence and some may be negative evidence. The current study explores the role of negative evidence in tasks requiring projections of novel properties. How does information about the presence or absence of a novel property in similar or dissimilar cases affect predictions? The literature reviewed above suggests three perspectives on negative evidence. The first, based on Bayesian conditional probability, is that negative evidence is uninformative. Whether or not all ravens are black, the number of non-black non-ravens is basically the same, so discovering one should not have a large impact on inferences. The second perspective derives from categorisation and word-learning research. Negative evidence may provide a foil or contrast to indicate a category. A white swan contrasts with a black raven at the species level, thus suggesting "raven" as the relevant category to use for generalisation. Finally, a third perspective presents a negative case as a competitor with positive evidence. A white swan and a black raven are alternative matches for some conclusion.

The experiments reported below address two issues. The first question is the evidential significance of negative evidence in property projection tasks. Does inclusion of negative evidence increase, decrease, or not affect people's willingness to project novel properties? This question is the central focus of Experiments 1a and 1b. The second question concerns the types or qualities of negative evidence. What sorts of negative cases provide stronger or weaker support for inductive inferences? Of particular interest are developmental differences in the treatment of negative evidence. If children and adults are using different strategies to make inductive inferences, then they should respond differently to negative evidence. This second question is the primary concern of Experiments 2 and 3.

The experiments in this study follow the conventions of the category-based induction literature. Inductive problems are referred to as "arguments" with presented evidence as premises and the unknown target as conclusion. Table 1 illustrates the types of arguments included in the experiments. As noted, the negative premises involve implicit rather than explicit negations. There were two reasons for this choice. First, other studies of triad induction that are the closest comparisons for the present research use implicit rather than explicit negations (Gelman & Markman, 1986; Sloutsky & Fisher, 2004; Springer, 1992). Second, identifying a premise with explicit negations is pragmatically odd and may introduce demand characteristics into the task. To present a white swan and state

TABLE 1
Examples of items used in experiments

Argument type	*Premises (evidence)*	*Conclusion(s)*	*Experiments using*
Single	Raven+*	Raven	1a, 1b, 2
Positive	Raven+, Swan−	Raven	1a, 1b
Basic Negative	Raven+, Swan−	Raven, Falcon, Dog	1a, 1b, 2, 3
Subordinate Negative	Raven+, Crow−	Raven	2
Superordinate Negative	Raven+, Cat−	Raven, Falcon, Dog	3
Kingdom Negative	Raven+, Tree−	Raven, Falcon, Dog	3

*The + or − represents the properties ascribed to the exemplar. For example, + might be "has a paxtin stomach", and − "has a fylate stomach". Examples given in this table are for illustration only. Actual properties and exemplars used in the experiments are listed in the appendices.

"This one is not black and is not a raven." is to specify the contrast classes relevant to the problem.

EXPERIMENT 1

Experiment 1 explores whether people are more or less likely to project properties when provided with negative evidence. The significance of negative evidence is assessed with respect to two comparison cases. The baseline argument is a single premise and matching conclusion. If told that one raven is black, what is the likelihood that people will predict that another raven is also black? Relative to single premise arguments, adding a negative premise may strengthen arguments (more likely to project), weaken arguments (less likely to project), or leave arguments unaffected. Adding a negative premise could affect arguments simply because more evidence is available, the kind of evidence may not matter. Thus, a second type of comparison case is an argument with matching and non-matching exemplars both ascribed the same property. This kind of argument provides two examples of individuals possessing the property in question (e.g., a black raven and a black swan have the same property). The initial question addressed in Experiment 1 is which kind of argument will result in the greatest likelihood of projection of properties from positive premises to conclusions.

Developmentally, the prediction is that negative evidence will be especially influential for young children. Research on categorisation and word learning suggests that young children do learn from comparing and contrasting cases (Namy & Gentner, 2002; Waxman & Klbanoff, 2000). Although there is some debate regarding whether young children use category-based processes to make inductions (Gelman, 2003; Sloutsky & Fisher, 2004), similarity-based approaches also predict a strong role for

negative evidence. The alternative, Bayesian prediction of weak effects of negative evidence is based on intuitions about property distributions and sampling probabilities. It seems plausible that children's judgements would be less affected by such factors than would adults'.

Method

Participants. A total of 17 young children ($M = 4;11$, range 4;4 – 5;8),16 older children ($M = 7;10$, range 7;5 – 8;11), and 22 adults participated in this experiment. All participants were drawn from the same medium-sized midwestern US city. Children were recruited from a birth registry database and from local preschools. Adults were recruited from undergraduate classes and participated for course credit. An approximately equal number of male and female children participated. Adults were predominately female.

Design. Participants evaluated arguments with three evidence types; Single, Positive, and Negative. Each trial involved property ascription to one (single) or two (positive and negative) exemplar(s). In the Single trials, participants were told that one exemplar (e.g., a rabbit) had a property. In the Positive and Negative trials participants were given information about a single exemplar (e.g., a rabbit) along with information about a second exemplar within the same superordinate level (e.g., beaver: beavers and rabbits are both mammals). In Positive trials the second exemplar was ascribed the same property as the first. In Negative trials the second exemplar was ascribed an alternative property.

For children, each evidence type was instantiated in four trials with biological property ascriptions and two trials with non-generalisable properties (18 trials in total). Biological properties were described as internal features of the exemplar (e.g., has a paxtin stomach). Non-generalisable properties refer to accidental features (e.g., has a scratch) or idiosyncratic attributes (e.g., is 3 years old) that would not be expected to generalise from one individual to another. Because of concern that adults might respond at ceiling when asked to project biological properties, these participants also received psychological property trials. For each evidence type adults responded to three biological trials, three psychological trials, and two non-generalisable trials (24 trials in total). Assignments of property, category, and evidence type were randomised across participants. The exemplar sets and properties used in this study can be found in Appendix A.

The response measure asked participants to predict whether the property would be true of the conclusion exemplar. The conclusion exemplar was always of the same species and involved the same property as the single instance (e.g., "Do you think this rabbit has a paxtin stomach?").

Participants also rated their confidence in their predictions on a 3-point likert scale (''very sure'', ''kinda sure'', ''not sure'').

Materials and procedure. All questions were accompanied by pictures and presented on a laptop computer. Children were interviewed individually in a quiet area at their preschool or at a research facility. They were told they were going to play a game that involved learning about different animals and answering some questions. Adults participated on individual computers in groups of up to 12. The experiment lasted approximately 15 minutes. Arguments were presented in random order, blocked with respect to argument type. Pairings of exemplars and properties were randomised across participants.

Scoring. Predictions (yes/no) and confidence ratings (very sure, kinda sure, not sure) were combined to yield composite projection scores. Projection scores ranged from -3 (no, very sure) to $+3$ (yes, very sure) with intermediate scores of -2 (no, kinda sure), -1 (no, not sure), 1 (yes, not sure), and 2 (yes, kinda sure).

Results

Mean projection scores are presented in Figure 1. Projection scores were analysed in a 3 (Age) $\times$ 3 (Evidence type) ANOVA. Because only adults rated psychological properties, these data were not included in the ANOVA and were analysed separately. There was a significant main effect of age, $F(2, 52) = 35.5$, $p < .001$, with adults giving higher scores than children (who did not differ, all pairwise comparisons, Tukey's HSD, $p < .05$). The main effect of evidence type was also significant, $F(2, 104) = 17.1$, $p < .001$. There was no age $\times$ evidence interaction, and simple effects revealed that the evidence effect held at each level of age, smallest $F(2, 104) = 3.4$, $p < .05$ for adults. Pairwise comparisons revealed that evidence conditions were ordered: Negative > Positive > Single.

A second level of analyses considered property-type differences. Adults were close to ceiling for ratings of biological items ($M = 2.4$) and near chance for non-generalisable items ($M = -.19$). This distribution contributed to evidence effects being weakest for adults in the ANOVA analysis. Psychological items provided a more sensitive measure of evidence effects. Projection scores for psychological items reproduced the Negative > Positive > Single ordering (Mean ratings: 2.1, 1.6, .95 respectively, all pairwise comparisons $p < .05$ two-tailed t-tests[1]). For all participants,

[1]For all pairwise t-tests and Wilcoxon tests reported, familywise error was controlled using Holm's procedure.

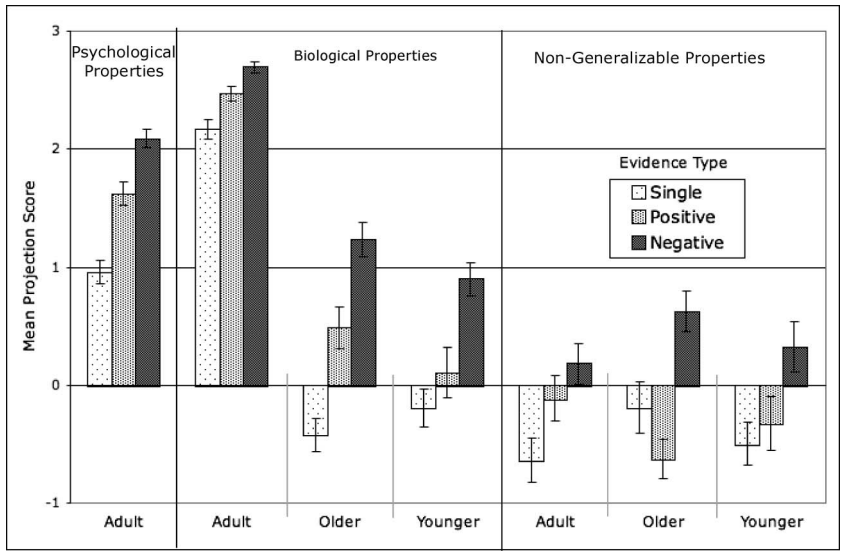

Figure 1. Proportion of positive projections for responses in Experiment 1a, across all evidence conditions and both property types. Bars represent one standard error.

evidence effects were less pronounced for non-generalisable properties. The test of the hypothesis that evidence conditions were ordered (monotonically decreasing) Negative > Positive > Single is provided by the Negative vs Single pairwise comparison (Marascuilo & Serlin, 1988). The order tests were significant for biological items, adults: $t(21) = 2.7$, older: $t(15) = 3.3$, younger: $t(16) = 3.2$, all $p < .05$. In no cases were the order tests significant for non-generalisable properties. Indeed, at no age were any pairwise comparisons among evidence conditions significant for non-generalisable properties.

The final set of analyses considered absolute rates of predictions and individual patterns. The prediction responses provide a clear comparison against chance (projection scores are most useful for relative condition comparisons). Each prediction question was a forced-choice between two options; thus chance performance would be .5. Table 2 provides the mean rates of prediction for Biological and Non-generalisable properties. Only when provided Negative evidence for Biological properties were children's rates of prediction greater than would be expected by chance, older: $T(15) = 110$, younger: $T(14) = 88.5$, both $p < .05$, two-tailed Wilcoxon tests. Adults reliably projected biological properties in all evidence conditions. Confirming the analyses presented above, in no cases did rates of prediction differ from chance for Non-generalisable properties.

TABLE 2
Mean proportions of Positive predictions, Experiment 1a

	Biological			*Non-generalisable*		
Age	*Single*	*Positive*	*Negative*	*Single*	*Positive*	*Negative*
Adult	0.92	0.97	1.00	0.39	0.52	0.59
Older	0.41	0.59	0.77	0.41	0.38	0.63
Younger	0.46	0.50	0.68	0.35	0.44	0.59

A basic test of individual patterns considered the number of participants who gave higher projection scores to one of the evidence conditions over the others. Consistent with the above analyses, many participants assigned higher scores to arguments with negative premises than to arguments with only a single premise. Considering only non-tied scores for biological properties, 13 of 15 younger children showed the pattern, as did 12 of 16 older, and 14 of 19 adults. Each of these frequencies differs from chance ($p < .05$, two-tailed sign test). In contrast, when properties were non-projectible, participants did not show a reliable preference for arguments with negative premises (9 of 16 younger, 8 of 12 older, and 10 of 18 adults, all $p > .05$, sign test). Adults and older children also gave higher ratings for Positive arguments than Single, with 13 of 17 and 12 of 16 non-tied cases in this direction (both $p < .05$). Fewer young children rated Positive arguments more highly. The pattern was shown by 8 of 15 younger children, a rate not significantly different from chance.

The general finding is that negative evidence supports inductive inferences. People are more likely to project novel properties when provided with negative cases. Moreover, participants made more confident projections for problems involving negative premises than for problems involving two positive premises. Before discussing the implications of these results, we report a second study that attempted to replicate the findings using a different response measure.

EXPERIMENT 1B

The task in Experiment 1a used a forced-choice response measure. Participants either endorsed or rejected ascription of a property to a novel exemplar in the conclusion. Although this measure of projection is common in studies of induction using triad tasks (Gelman & Markman, 1986; Sloutsky & Fisher, 2004) it is potentially problematic in the current study. The concern is that presenting the induction question as a forced choice may have encouraged participants to adopt a matching strategy for evaluating premises. Participants may not have been judging how well the

evidence supports a conclusion, but rather how similar the conclusion was to the premises. A related concern is that the negative evidence condition provides a model or basis of comparison for both a "yes" (positive exemplar) and "no" (negative exemplar) response. The other evidence conditions did not have this clear mapping between exemplars and response options.

To address concerns about the forced-choice measure used in Experiment 1a, Experiment 1b used a confidence rating to assess inductive projections. In Experiment 1b, participants were presented with a conclusion (e.g., "This animal has paxtin in its blood") and asked how confident they were that the conclusion is correct. This response measure involved only a single property (the one ascribed to single and positive exemplars). As the conclusion already ascribes one of the properties to the novel exemplar, the task requires more than simply picking which premise exemplar is the best match. This structure also focuses more clearly on the question of the evidential value of the premises. The question is not which property the conclusion will have, but rather how well the evidence supports a particular ascription. An additional benefit is that this response measure was expected to reduce the ceiling effect for adult participants.

Method

Participants. A total of 18 young children ($M = 4;9$, range 4;3 – 5;7), 18 older children ($M = 7;7$, range 7;0 – 8;8), and 18 adults participated in this experiment. All participants were drawn from the same population and given the same reimbursement as in Experiment 1a. No individual participated in other experiments reported in this study.

Design. Participants responded to 18 items. There were six items from each of three evidence types: Single, Positive, and Negative. The properties and categories were the same as those used in Experiment 1a; for each evidence type participants received four biological items and two with non-generalisable properties. Rather than predicting whether the conclusion would have the property (as in Experiment 1a) participants were asked to rate their confidence that a conclusion exemplar had the same property as the species-matched exemplar. After given evidence (e.g., in the single case, "This rabbit has a paxtin stomach") participants were asked, "How sure are you that this other [exemplar] has a paxtin stomach?" There were five options to choose from: "not very sure", "somewhat sure"; "kinda sure"; "quite sure", and "very sure". There was no follow-up question. With the exception of the response measure (and exclusion of psychological properties for adults), Experiment 1b followed the same design as Experiment 1a.

Materials and procedure. The materials and procedures were the same as in Experiment 1a.

Scoring. Reponses were scored from ascending order of certainty: "not very sure" = 1, "somewhat sure" = 2, "kinda sure" = 3, "quite sure" = 4, and "very sure" = 5.

Results

Mean confidence ratings are presented in Figure 2. These ratings were analysed in an ANOVA with evidence type (Single, Positive, Negative) a within-subjects variable and age (Adults, Older, Younger) between subjects. The analysis revealed an effect of age $F(2, 51) = 6.6$, $p < .01$. Adults' confidence ratings were significantly higher than younger children, Tukey's HSD, $p < .05$. There was also a main effect of evidence type $F(2, 102) = 7.3$, $p < .01$. Confidence ratings were higher for Negative than for Single evidence, Tukey's HSD, $p < .05$. The main effects were conditioned by an age × evidence type interaction $F(4, 102) = 2.8$, $p < .05$. There was no effect of evidence manipulation for adults, $F(2, 102) = 0.4$. Type of evidence did significantly affect younger and older children's confidence ratings, $F(2, 102) = 8.6$ $p < .001$, and 3.8 $p < .05$, respectively.

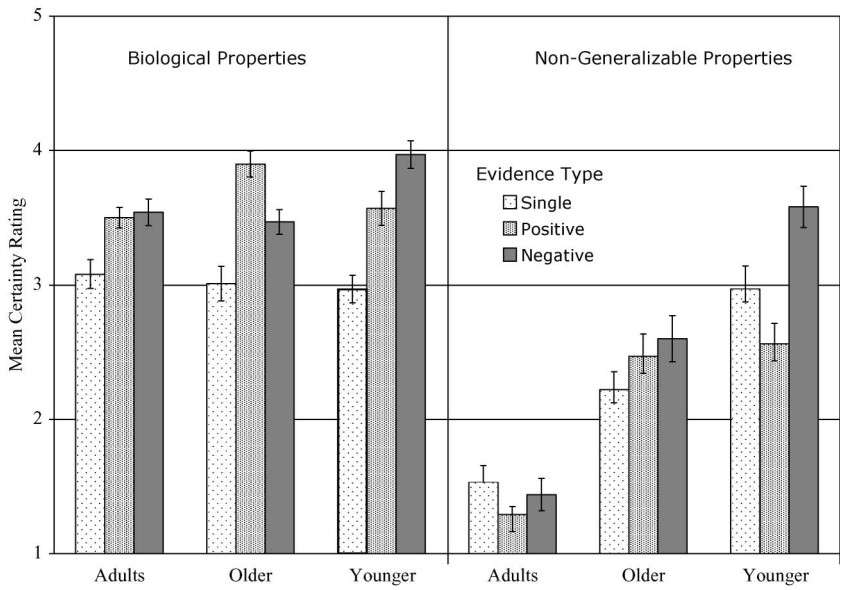

Figure 2. Mean certainty ratings for responses in Experiment 1b across all evidence conditions and both property types. Bars represent one standard error.

The preceding analysis considered ratings of both generalisable (biological) and non-generalisable properties. However, evidence effects were predicted to be different for the two kinds of properties. To simplify the analysis, data from each age group were considered separately. For younger children there was a main effect of property type. Confidence was greater for biological than non-generalisable properties, $F(1, 17) = 10.5$, $p < .005$. The main effect of evidence was also significant, $F(2, 34) = 5.1$, $p < .05$. Confidence was greater for arguments involving negative evidence than those involving single or positive evidence, (which did not differ Tukey's HSD $p < .05$). However, the effect of evidence was only significant for biological properties, $F(2, 34) = 6.6$, $p < .005$; $F(2, 34) = 2.9$, *ns*, for non-generalisable properties. Older children showed a similar pattern: greater confidence for biological properties, $F(1, 17) = 37.1$, $p < .001$, and a main effect of evidence, $F(2, 34) = 4.3$, $p < .05$. However, for older children, it was arguments with positive premises that were rated strongest (versus single arguments, no other comparisons were significant, Tukey's HSD, $p < .05$). Evidence levels were only different for biological properties. The pattern for adults was similar: significant effect of evidence for biological properties $F(2, 34) = 3.5$, $p < .05$, but not non-generalisable $F(2, 34) = 0.3$.

An additional set of analyses looked at absolute rates of projection. To measure absolute patterns, the mid-level of certainty ("kinda sure" = 3) was taken as the chance level of responding. Table 3 provides the mean ratings for both property types. At all ages, ratings were significantly higher than the chance level of certainty for biological properties given positive and negative evidence (all *p*s $< .05$, two-tailed Wilcoxon tests). For all evidence types, adult judgements were below chance levels for non-generalisable properties—single $T(17) = -5.7$, positive $T(17) = -13.2$, and negative $T(17) = -5.9$, all $p < .001$—while older children were also below chance in their responses to non-generalisable properties in the single evidence condition, $T(17) = -2.9$, $p = .01$. In all other evidence conditions children responded at chance levels for non-generalisable properties.

The small number of items precludes sensitive tests of individual patterns of responding. One way to address the question of whether patterns at the

TABLE 3
Mean certainty ratings, Experiment 1b

	Biological			*Non-generalisable*		
Age	*Confirm*	*Contrast*	*Single*	*Confirm*	*Contrast*	*Single*
Adult	3.5	3.54	3.08	1.29	1.44	1.55
Older	3.9	3.48	3.01	2.47	2.6	2.22
Younger	3.57	3.97	2.97	2.56	3.58	2.97

group level held for individuals is to assess the rank orderings of the evidence conditions. As in Experiment 1a, this analysis counts the number of participants who rated negative-premise arguments higher than single-premise arguments. Most participants did rate negative-premise arguments higher when arguments involved projectible (biological) properties. Of the non-tied scores, 13 of 16 younger children rated negative arguments higher, as did 12 of 15 older children, and 11 of 15 adults. These frequencies are significantly different from what would be expected by chance ($p < .05$, sign test), except for adults ($p = .06$). There was no consistent ordering when arguments involved non-generalisable properties: Younger, 8 of 13; Older, 10 of 16; adults 4 of 9 (all $p > .05$, sign test). Only older children consistently rated positive arguments more highly than single. Of 15 non-tied rankings, 13 showed this pattern ($p < .05$, sign test), while the frequencies for younger children and adults did not differ from chance (11 of 16 and 11 of 17, respectively).

The results of Experiment 1b are generally consistent with those of Experiment 1a. Across both experiments, negative evidence increased rates and confidence of projection compared to single-premise arguments. At least for the younger children in the experiments, negative evidence may have been more significant than additional positive evidence. In Experiment 1a, young children's projection scores were above chance only when provided negative evidence. A significant proportion of young children rated negative-premise arguments more highly than single-premise arguments, but the rate was not higher for positive-premise arguments. In Experiment 1b, young children were more confident given negative evidence than when given additional positive evidence. The relative significance of positive and negative evidence was less clear for adults and older children. In general, both types of additional evidence were significant, but one was not consistently more influential than the other.

The results of Experiments 1a and 1b are inconsistent with the Bayesian predictions as described in the Introduction. If negative evidence is equally likely whether or not the conclusion is true (the chance of encountering a non-black non-raven is the same whether all ravens are black or not) then negative evidence should have little impact on assessments of the conclusion. A plausible response is that this prediction mis-states the likelihoods. In particular, the Bayesian prediction is based on the assumption that premises are being randomly selected from some large population (e.g., all animals). The results of Experiments 1a and 1b suggest that this is not how participants (particularly young children) understood the selection of evidence. We will return to a discussion of sampling and the origins of premises in the General Discussion. At this point, it is sufficient to note that the category-based hypothesis, that negative evidence is treated as a foil lying just outside a category boundary, implies that the selection of evidence is not random.

Both category- and similarity-based accounts are consistent with the findings that negative evidence has a strong influence on inductive projections. Experiments 1a and 1b provide some modest challenges to similarity-based accounts. The similarity-based hypothesis is that participants are comparing the premise exemplars with the conclusion. Negative evidence turns the task into a forced-choice judgement (is the conclusion more similar to the positive or negative exemplar?), while positive-only arguments involve an assessment of absolute similarity (is the conclusion similar enough to an exemplar to warrant projection?). Under such an account, the property being projected may not matter. Yet participants did not show evidence effects for projections of non-generalisable properties. Moreover, the relative strength of positive and negative premises was the same in both a forced-choice task (Experiment 1a) and in a confidence-rating task (Experiment 1b). These results provide some suggestion that participants were not simply making a forced-choice judgement about which exemplar best matched the target in arguments involving negative premises. Clearly, though, these results are not definitive. A stronger test of the category- and similarity-based hypotheses comes from a consideration of different types of negative evidence. Both perspectives suggest that negative evidence may provide strong support for induction. However, predictions differ regarding the features that make negative evidence strong. These predictions are described and evaluated in Experiments 2 and 3.

EXPERIMENT 2

On a similarity-based model, positive and negative premises are competitors. If people evaluate arguments with both positive and negative premises by deciding which premise is most similar to the conclusion, then very dissimilar negative premises should make the strongest argument. It is easier to determine that a Lion is more similar to a Tiger when the alternative is a Bee than when the alternative is a Bear. This intuition is formalised in models that present choice or inductive projection (Sloutsky & Fisher, 2004) as a function of the ratio of the similarities of the positive and negative options to the conclusion. Thus the prediction is that the argument ''Lions have X, Bees have Y, therefore Tigers have X'' will be stronger than the argument ''Lions have X, Bears have Y, therefore Tigers have X''. The less similar the negative premise is to the conclusion, the less well it competes with the positive premise, and the easier it is to match the positive premise and the conclusion.

The category-based account of negative evidence is somewhat more complicated. The hypothesis is that negative evidence supports arguments by indicating or making relevant (Medin, Coley, Storms, & Hayes, 2003) a category that includes the positive premise and the conclusion. Research on

word and category learning has generally held that a contrast is most helpful when it falls just outside the boundary to be learned (Au & Markman, 1987; Houtz et al., 1973). With respect to the stimuli used in the current study, an informative contrast or foil will be a non-instance that is a member of the category immediately superordinate to the category containing the positive premises and conclusion. Continuing the example from the previous paragraph, Bear is a stronger contrast than is Bee because Bears share membership with Lions and Tigers in a more immediate superordinate category. Put slightly differently, information that Bears lack the property rules out more alternatives than information that Bees lack the property. The property in question is not true of all mammals, or all carnivores, etc. Previous research has shown that the number of alternative possibilities is negatively related to argument strength (McDonald, Samuels, & Rispoli, 1996). The original similarity-coverage model (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990) makes a similar prediction. Take the covering category in an argument with negative premises to be the category that includes all positive premises and the conclusion, but excludes all negative premises. The more general the category, the less well the positive premises will cover (represent) the category, and the weaker the argument. The construction of categories, or relations between premises and conclusions, can be quite complex (see Medin et al., 2003), however the basic intuition of the category-based account, in contrast to the similarity-based account, is that negative evidence may be too dissimilar to the relevant category to be informative. A very similar contrasting case is the most informative.

The results of Experiments 1a and 1b suggest that negative evidence can have a strong influence on inductive projections. The goal of Experiments 2 and 3 is to begin to explore the nature of that influence; what makes a strong piece of negative evidence? The strategy is to manipulate the similarity and shared membership of negative premises with respect to conclusions and positive premises. We consider two hypotheses about the effects on inductive projections of variations in negative evidence. The similarity-based prediction is that inductive projections should be inversely related to similarity of negative premises: As negative premises become less similar to conclusions, projections should increase. The category-based prediction is that projections should be positively related to negative premise similarity: The more similar the negative premise is to the conclusion (more categories in common) the stronger the argument.[2] Experiments 1a and 1b asked

[2]To a point. To be a negative premise, the exemplar in the premise must be a non-instance of the most narrow category containing the positive premises and the conclusion. A very similar negative premise becomes a counter-example (instance of the category that lacks the property) and should reduce argument strength.

whether negative evidence was more or less effective than single or additional positive evidence. The focus of Experiment 2 is not whether negative evidence supports inductive inferences, but rather which kinds of negative evidence provide the most support.

Method

Participants. A total of 15 young children ($M = 4$; 10, range 4;3 – 5;5), 17 older children ($M = 8$;1, range 7;7 – 8;9), and 22 adult undergraduates participated in this experiment. All participants were drawn from the same population as other experiments reported here. Participants did not take part in any of the other experiments.

Design. Participants responded to 16 trials. There were four trials each from four different evidence types—Single, and three types of negative evidence: Subordinate level, Basic level, and Superordinate level. The single-evidence trials were presented in the same way as in Experiments 1a and 1b (e.g., "This rabbit has a paxtin stomach."). All other conditions involved a single (positive) premise then one of three types of negative evidence. The basic-level trials were similar to the negative-evidence conditions reported in Experiments 1a and 1b, with the negative evidence involving an exemplar from a different basic-level category (e.g., "This beaver has a fylate stomach."). Negative premises for superordinate-level trials were drawn from distinct lifeforms (e.g., bird or reptile for mammal conclusions). For subordinate-level trials, the positive and negative exemplars were given distinctive sub-species labels (e.g., "jackrabbit" and "cotton-tail rabbit"). The projection question was the same as in Experiment 1b; participants were asked to judge the likelihood a property was true of the target. The conclusion was always an instance of the exemplar named in the positive premise. For example in the single, basic-level, and superordinate-level trials participants were asked to make projections to another "rabbit"; in the subordinate-level condition they were asked about another "jackrabbit". Assignment of properties and conclusion exemplars to evidence conditions was randomised across subjects. A list of the exemplars used in this experiment is available in Appendix B.

Materials and procedures. The procedures were identical to those used in Experiments 1a and 1b. The only modification was the incorporation of a new set of pictures to reflect the change in stimuli.

Scoring. Responses were scored the same way as in Experiment 1b, from ascending order of confidence: "not very sure" = 1, "somewhat sure" = 2, "kinda sure" = 3, "quite sure" = 4, and "very sure" = 5.

Results and discussion

Mean confidence ratings are presented in Figure 3. The first analysis involved an ANOVA with Evidence type (Single, Superordinate level, Basic level, Subordinate level) as the within-subjects variable and Age (Adults, Older children, Younger children) as the between-subjects variable. There was no main effect of age, nor was the interaction between age and evidence type significant. The analysis revealed a main effect of Evidence type $F(3, 144) = 3.7$, $p < .05$. Confidence ratings were significantly lower in the single-evidence case than the subordinate-level case (Tukey's HSD, $p < .05$). This result provides some support for the category-based hypothesis. The negative premises providing the most specific contrast were the only ones to yield higher confidence ratings than single-premise arguments. At the same time, pairwise comparisons failed to demonstrate significant differences between the various negative evidence conditions.

A planned comparison revealed that, averaging across subordinate, basic, and superordinate levels, arguments with negative premises were significantly stronger than arguments with only a single premise. Only for adults did some negative evidence produce reliably higher confidence ratings than single evidence, $F(1, 19) = 4.8$, $p < .05$. Although the trends for both younger and older children are in the same direction, evidence-type differences did not reach statistical significance.

The results of Experiment 2 did not provide conclusive support for any of the hypotheses about negative evidence. The only reliable finding was that

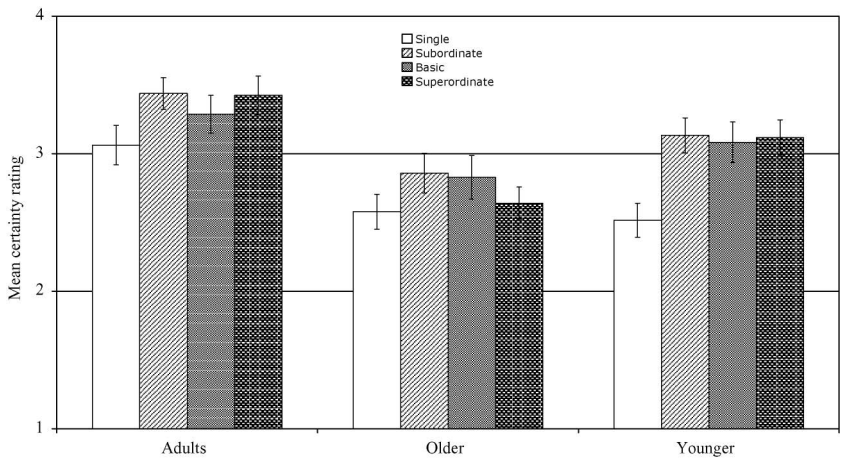

Figure 3. Mean certainty ratings for responses in Experiment 2 across evidence conditions. Bars represent one standard error.

negative evidence at the most specific level of contrast produced higher confidence ratings than single premises alone. This result is consistent with the category-based prediction. However, the data do not support conclusions about differences between the various negative evidence conditions. In this regard, neither the category-based nor the similarity-based accounts received support. While there were no consistent differences between the levels of negative evidence presented in the experiment, the equivocal pattern of results indicates there was substantial unexplained variance in the argument ratings.

The primary conclusion from Experiment 2 was negative; there was no evidence that the kind of contrasting premise affected generalisations. This kind of negative conclusion is compelling only to the degree that the measure is a powerful or sensitive one. There are several features of the task in Experiment 2 that might have limited the ability to detect differences between negative premises. Most critically, the positive premises were always very similar to (shared subordinate category with) the conclusions. Perhaps the high degree of match between positive premises and conclusions overwhelmed any influence of variation in negative premises. A second limitation was that the range of negative premises was relatively narrow. All the exemplars in the arguments were from the same kingdom classification. Arguments with a wider range of premises would have a greater chance of revealing consistent effects of levels of negative evidence on projections. Finally, the inclusion of single-premise arguments in the design may have overshadowed the differences among arguments with negative premises. The most salient feature of the items may have been the difference between those with one premise and those with two premises. Participants may not have been attending to the variations in the negative arguments, rather simply tracking the difference between arguments with less evidence and those with more.

EXPERIMENT 3

The goal of Experiment 3 was to assess whether the finding of a lack of distinction between types of negative evidence would be replicated with a more extreme set of examples. This task also provides a more robust test of the category- and similarity-based hypotheses. The central change from Experiment 2 was manipulation of the relation between the positive premise and the conclusion. By varying this relation, the effect of negative premises may be more evident. The similarity-based prediction is that the relation between positive premise and conclusion, and the relation between negative premise and conclusion, are independent effects on argument strength. As Sim(positive premise, conclusion) increases, argument strength

increases. As Sim(negative premise, conclusion) decreases, argument strength increases.[3]

The category-based prediction is that the relations between the premises and conclusions are jointly considered; there will be an interaction between conclusion level and negative premise type. Reasoners use premises to identify a category of things that possess the property in question. To the extent that the various premises (e.g., positive and negative) indicate the same category, the argument will be strong. The mechanisms whereby premises indicate, or make relevant, a category are complex and variable (Medin et al., 2003). For the purposes of the current study the hypothesis is that a positive premise indicates the lowest-level category containing the premise and the conclusion (see Osherson et al., 1990). The positive premise of "raven" and conclusion of "swan" indicates the category "bird". A negative premise that contrasts at the same level (e.g., "bat") will provide the greatest support to the argument. Negative premises contrasting at other levels (e.g., more general "animal", or less general "sparrow") will have weaker effects because they do not clearly indicate the same category.

Method

Participants. A total of 16 young children ($M = 5;1$, range 4;6 – 5;8), 16 older children ($M = 7;8$, range 7;4 – 8;6), and 20 adult undergraduates participated in this experiment. All participants were drawn from the same population as other experiments reported here. Participants did not take part in any of the other experiments.

Design. Participants rated nine sets of three arguments. Each argument included a positive premise (e.g., "This {red} wolf has a paxtin stomach."). Both conclusion and negative premise types are defined relative to the positive premise of the argument. There were three arguments each from one of three negative evidence conditions: basic, superordinate, or kingdom contrast (e.g., horse, fish, and rose, respectively). For each set of arguments there were three conclusion targets: basic, superordinate, and kingdom match (e.g., timber wolf, fox, and bird, respectively). The conclusion match

[3]The effects are not totally independent, as the magnitude of the contribution of negative evidence may change depending on the similarity of positive premise and conclusion. If the positive premise is quite similar to the conclusion there is little "room" for additional negative evidence to influence predictions (judgements are close to ceiling). However the relative significance of different kinds of negative evidence (i.e., very similar, very dissimilar) should not change. The focus of Experiment 3 is the relative significance of different kinds of negative evidence (e.g., is less similar always better?) not the absolute impact of negative evidence on predictions for different conclusions. Thanks to Vladimir Sloutsky for noting this point.

refers to the most specific category containing both the positive premise and the conclusion (e.g., red wolf and timber wolf are both wolves). The negative contrast refers to the level at which the positive and negative premise would not share category membership (e.g., a wolf and a horse differ in basic-level category but share a superordinate: mammal). For simplicity, negative premises will be referred to by their level of contrast. Although the stimuli are described in terms of taxonomic categories, the structure could also be characterised by similarity. There were three levels of conclusion similarity, defined relative to positive premise: high, medium, and low. There were also three levels of negative premise: high, medium, and low. From the similarity-based perspective it is the relation between negative premise and conclusion that is critical. Stimuli for negative premises were selected such that the high, medium, low designation held for similarity to both positive premise and to conclusion. For example, a rose (kingdom contrast) is lower in similarity to both a wolf (positive premise) and fox (superordinate conclusion), than is a horse (basic contrast). The similarity structure of the stimuli was confirmed in a pre-test (see below).

The response measure was a forced choice followed by confidence rating as used in Experiment 1a. This measure yielded the strongest effect of negative evidence. As Experiment 3 did not involve Single or Positive arguments, concerns about the matching demands of the forced-choice method were not relevant.

Procedure. Argument sets were presented in random order. Participants rated each of the three conclusions for a set (in random order) before seeing the next set of premises and conclusions. In other respects the procedure was identical to that of Experiment 1a.

Similarity pre-test. The logic of the design required that the relative similarities of contrasts be constant across conclusion types. Basic contrasts should be more similar to all conclusions than are superordinate contrasts, which should be more similar to all conclusions than are kingdom contrasts. A total of 15 adults participated in a similarity-rating task to check that the stimuli used in the experiment conformed to these constraints. Each participant rated the similarity for all contrast × conclusion pairs used in the induction task. The rating task used a 9-point scale. Similarity ratings were analysed in an ANOVA with conclusion and contrast levels as within-subjects factors. There was a main effect of conclusion, $F(2, 28) = 13.5$, with positive premises less similar to kingdom conclusion than to basic or superordinate (collapsing across contrast type, Tukey's HSD, $p < .05$). The main effect of contrast level was significant, $F(2, 28) = 54.6$, $p < .001$. Across conclusions, basic-level contrasts received higher similarity ratings than did superordinate contrasts, and both received higher ratings than kingdom

contrasts (all comparisons $p < .05$, Tukey's HSD). The ANOVA also revealed a significant conclusion by contrast interaction, $F(4, 56) = 37.0$, $p < .001$. Pairwise comparisons showed that the contrast differences were as predicted for basic- and superordinate-level conclusions. For kingdom-level conclusions, basic and superordinate contrasts did not differ significantly (although both were different from kingdom, all comparisons $p < .05$, Tukey's HSD). As the main predictions for kingdom-level conclusions concerned kingdom contrasts, the interaction in the similarity ratings was not a major flaw in the stimuli.

Results

Figure 4 shows the mean projection scores at each level of conclusion and contrast. The central question was whether the contrast presented would have a consistent effect across all levels of conclusion, or whether the variables of contrast and conclusion would interact. Separate ANOVAs for each age group were conducted with contrast level and conclusion level as within-subjects factors. Adults showed no consistent main effect of contrast,

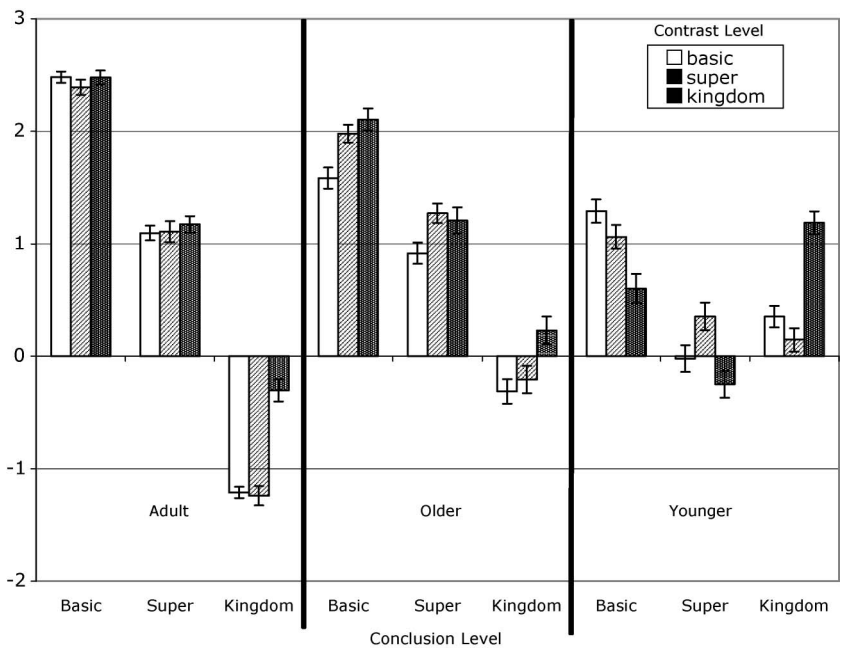

Figure 4. Mean projection scores for responses in Experiment 3 across all levels of contrast and conclusions conditions. Bars represent one standard error.

but there was a main effect of conclusion, $F(2, 30) = 58.6$, $p < .001$. As predicted on a similarity-based account, the similarity between positive premise and conclusion was positively related to probability of projection. Adults were more likely to project to a basic-level conclusion than a superordinate level, and more likely to project to a superordinate conclusion than a kingdom (all comparisons $p < .05$, Tukey's HSD). The conclusion × contrast interaction did not meet the established criteria for statistical significance, $F(4, 60) = 2.5$, $p = .056$. Older children showed a pattern of results similar to adults. The only significant effect was the main effect of conclusion, $F(2, 30) = 7.9$, $p < .01$. Older children were also more likely to project to the more similar conclusions, although only the pairwise comparison between basic and kingdom conclusions was statistically significant ($p < .05$, Tukey's HSD). The general finding is a main effect of positive premise–conclusion similarity on older children's and adult's inductive projections.

As is apparent from Figure 4, younger children showed a very different pattern of responding from older children or adults. The ANOVA revealed no main effects for either conclusion or contrast. However, the interaction was significant, $F(4, 60) = 5.5$, $p < .001$. Thus young children did not reliably make more projections to more similar conclusions. Instead, the type of negative premise determined which conclusion was most likely to be ascribed the property of the positive premise.

The similarity- and category-based hypotheses make different predictions about which contrast will be the most effective. Given the construction of the stimuli, the kingdom-level contrasting premise was always the most dissimilar to the conclusion. Thus the similarity-based prediction is that arguments with the kingdom contrasts will yield the strongest projections. The category hypothesis is that a negative premise contrasting at the same level as the conclusion will be the strongest. Three post-hoc comparisons (one at each level of conclusion) tested these alternative predictions. In each case, the contrasts predicted to be strongest and weakest on the category hypothesis were compared. For kingdom-level conclusions, the similarity and category predictions are the same: The kingdom contrasts will be strongest, and the basic contrasts will be weakest. For the other two conclusions, the predictions are opposed. The similarity prediction is that kingdom contrasts will be strongest; the category prediction is that kingdom contrast will be weakest given basic- and superordinate-level conclusions (with basic and superordinate contrasts strongest, respectively). Given that the experiment was designed to allow a test of these alternative predictions, post-hoc comparisons were conducted for all three age groups.

As expected from the non-significant interactions, none of the comparisons was statistically significant for older children. This pattern is inconsistent with both the similarity and the category hypotheses. Instead, the

data suggest that older children were not attending to the negative premise, but basing their predictions solely on the relation between the conclusion and the positive premise. There was a single significant contrast difference for adults: projections to kingdom conclusion were stronger given kingdom than basic contrasts, $F(1, 60) = 11.3$, $p < .005$. This result is consistent with both the similarity and the category hypotheses. The comparisons for young children revealed the clearest discrimination among hypotheses. In each of the three comparisons, young children showed the pattern predicted by the category hypothesis. For basic-level conclusions, basic-level contrasts were stronger than kingdom contrasts, $F(1, 60) = 4.4$, $p < .05$; for superordinate-level conclusions, superordinate contrasts were stronger than kingdom, $F(1, 60) = 4.4$, $p < .05$; and for kingdom-level conclusions, kingdom contrasts were stronger than basic, $F(1, 60) = 7.3$, $p < .01$.

The final set of analyses considered the number of participants whose ratings were consistent with the category-based predictions. A consistent rating was defined as providing a higher rating for arguments predicted to be strongest than for arguments predicted to be weakest for each level of conclusion. As the similarity- and category-based predictions were the same for kingdom-level conclusions, these items were not included in the analysis. Ignoring ties, the chance probability that an individual participant would rate the predicted argument higher is .5. The probability of responding as predicted at both levels of conclusion (subordinate and basic) is .25. Of the 16 younger children, 12 had non-tied scores and could be included in the analysis. Of the 12, 8 showed the predicted rating pattern for both levels of conclusion. This frequency is different from what would be expected by chance ($p < .005$, Binomial Theorem). A total of 11 adults and 12 older children had non-tied scores. In each case, however, only two participants at each age showed the predicted pattern.

Discussion

For older children and adults, the results of Experiment 3 replicated the non-significant findings of Experiment 2; the level of negative evidence did not have a consistent impact on argument ratings. The nature of the positive premise was the most significant influence on inductive projections. The more similar the positive premise and the conclusion, the more likely adults and older children were to project properties. The one reliable influence of variations in negative evidence for adults was restricted to cases where the positive premise and conclusion were dissimilar. In such cases it was the least similar negative evidence that produced the strongest arguments. The results are not informative about the alternative predictions: The category- and similarity-based predictions were the same for the one condition in which variation in negative evidence did affect adults' judgements.

Unlike the older participants, younger children showed a consistent response to variation in negative evidence. Negative premises at the same level of generality as conclusions produced the strongest arguments. Critically, negative premises that were very dissimilar to the conclusions (that shared membership in only a very general superordinate category) often produced fewer/weaker projections than arguments involving more similar negative premises. These results are inconsistent with a model in which young children are directly comparing the relative similarities of positive and negative premises to the conclusions. In contrast to the frequent finding in the literature, it is older children and adults who appeared to be using similarity-based strategies. Younger children's performance suggests they were using both positive and negative evidence to identify a category for projection.

GENERAL DISCUSSION

Overall, the results of Experiments 1–3 suggest that negative evidence increases the likelihood that adults and young school-aged children will project properties to a novel exemplar. However, the older participants in the experiments were less influenced by negative evidence than were the younger (preschool-aged) participants. At all ages, adding a negative premise to a single (positive) premise argument produced more and stronger projections. For adults and older children this result may reflect the amount of evidence available. Arguments with more premises are stronger than arguments with fewer. Providing negative premises had roughly the same effect as adding additional (dissimilar) positive premises (in Experiment 1b). At least as far as the experiments in the current study (Experiments 2 and 3), any piece of negative evidence contributed about equally to older participant's judgements of argument strength. Older participants were largely insensitive to variations in the quality of negative evidence. In contrast, younger participants did vary their projections according to the type of negative evidence included in arguments. This result indicates that younger children were more influenced by negative evidence than were older children and adults. Both the presence and the quality of negative evidence mattered for preschool-aged children, while for older participants it was only the amount of negative evidence (presence/absence) that had a consistent effect on inductive projections.

Negative evidence had more influence on inductive projections at all ages than was predicted by the Bayesian analysis. This result does not imply that Bayesian models are poor accounts of human inference. Rather, this study points to the need to more carefully examine participants' priors and assumptions. Common methods for assessing inductive inference (like those used in Experiments 1–3) typically do not provide all the information that

might be relevant to drawing a conclusion. For example, participants are often not told the base rates of exemplars (how many ravens are there?) or properties (how many black things?). McKenzie and Mikkelsen (2007) have shown that when positive exemplars and properties are known to be common, and negative values are known to be rare, negative evidence is treated as informative. It is not clear why participants in the current experiment should assign priors in this way. It is not implausible that there would be developmental differences in assumptions about base rates. One direction for future research on inductive inference is studying how evidence affects people's beliefs about the frequencies of the exemplars and properties encountered.

Treatment of evidence also depends on beliefs about sampling strategy. It is rare for participants in induction studies to be told how the evidence was collected. In the triad task, no one ever explains how or why the examples were selected. In the absence of information about sampling strategy it is simply not possible to judge the evidential significance of some premise. The rarity effect is due to intuitions about sampling (rare evidence should show up less frequently). A piece of negative evidence selected at random has a very different implication from a piece of negative evidence selected by a helpful teacher as an illustration of a category boundary. Especially when the type of negative evidence varied (in Experiments 2 and 3), participants likely formed some (implicit) ideas about how the evidence was generated. Young children were responding to negative premises as if they were good foils. If evidence is randomly selected, then it seems unlikely that negative premises would represent instances just outside the category boundary. If evidence is being selected to be informative, to indicate category boundaries, then a good foil is to be expected. The insensitivity to the type of negative evidence shown by adults and older children may reflect some intuitions that the evidence was selected at random. One way to test this hypothesis is to explicitly provide information about sampling. Perhaps people's use of negative evidence would change depending on whether they thought the premises were carefully selected to be informative or were randomly generated from some large population.

One contribution of the current study is to suggest avenues for future research. A more direct contribution concerns debates about young children's inductive projections. Young children are insensitive to many of the inductive principles that adults use (Gutheil & Gelman, 1997; Lopez et al., 1992). A basic finding in the literature on category-based induction is that adults generate inferences by constructing a covering category that includes the (positive) premises and conclusion. The strength of the argument, the likelihood of making an inductive projection, depends on the homogeneity of the covering category, and the degree to which the evidence covers or represents that category (Osherson et al., 1990).

Developmental findings indicate children younger than 5 years old do not integrate information in inductive arguments to form hypotheses about covering categories (Lopez et al., 1992). The general argument has been that young children rely on similarity matching while older children and adults tend to base their inferences on representations of categories (Sloutsky & Fisher, 2004). The findings from Experiment 3 challenge this view and provide evidence that young children do use premises to identify a category to guide their inductive inferences.

The core of a psychological account of inductive inference is a description of the processes people use to construct a connection between known cases and novel ones. Told that object A has a property, one's conclusion about object B depends on the salience of some relation between A and B. Debates concern just which kinds of processes people use to establish and evaluate relations. A plausible developmental hypothesis is that relevant relations are initially based on similarity. The significance of a known case depends on its similarity to the novel case.

Experiment 3 in the current study suggests that a straightforward account of similarity relations is insufficient to characterise young children's inductive inferences. The relative similarity of two pieces of evidence did not predict their impact on young children's likelihood of making an inductive projection. Information about positive and negative premises was not combined in a similarity ratio. Rather, the contribution of premises is more aptly characterised as evidential. Certain combinations of premises indicated, suggested, or made relevant (Medin et al., 2003) connections to conclusions. The mechanisms that generate relevant connections can be complex and difficult to specify. In this case the demonstration was that premises that contrast at an appropriate level of generality provide greater support for inductive projections than premises that are very dissimilar. Whether or how young children differ from adults and older children in the kinds of processes they use to establish relevant connections for inductions remains an open question. The contribution of the current study is to demonstrate that young children use mechanisms of category formation, at least in addition to judgements of similarity.

Many of the principles of inductive inference that have been explored in the literature are based on statistical principles. Whether larger, more diverse, and more surprising sets of evidence increase inductive confidence (Gutheil & Gelman, 1997; Lo et al., 2002; Lopez et al., 1992) are questions that make sense against a background of assumptions about probability and random sampling. Young children may not share adult intuitions about base rates or the sources of evidence. Their methods of evaluating evidence may rely heavily on assumptions of helpfulness and conversational implicature (Siegal & Surian, 2004). Thus young children's patterns of inductions may not appear to involve assessment of evidence and evaluations of hypotheses.

Of course, such results are negative evidence. That young children do not use some strategies of evidence evaluation does not mean they use no strategies, or use only similarity-based comparisons. In the experiments reported in the current study, children's use of negative evidence provides positive evidence for category-based strategies in inductive inference.

Manuscript received 19 April 2006
Revised manuscript received 7 September 2006
First published online 23 May 2007

REFERENCES

Au, T. K., & Laframboise, D. E. (1990). Acquiring colour names via linguistic contrast: The influence of contrasting terms. *Child Development*, *61*, 1808–1823.

Au, T. K., & Markman, E. M. (1987). Acquiring word meanings via linguistic contrast. *Cognitive Development*, *2*, 217–236.

Billman, D., & Davilla, D. (2001). Consistent contrast aids concept learning. *Memory & Cognition*, *29*, 1022–1035.

Eells, E. (1982). *Rational decision and causality*. Cambridge, UK: Cambridge University Press.

Evans, J. St. B. T., & Handley, S. J. (1999). The role of negation in conditional inference. *The Quarterly Journal of Experimental Psychology*, *52*A, 739–769.

Feldman, S. S. (1972). Children's understanding of negation as a logical operation. *Genetic Psychology Monographs*, *85*, 3–49.

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

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

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.

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

Hempel, C. G. (1945). Studies in the logic of confirmation. *Mind*, *54*, 1–26.

Houtz, J. C., Moore, J. W., & Davis, J. K. (1973). Effects of different types of positive and negative instances in learning "nondimensioned" concepts. *Journal of Educational Psychology*, *64*, 206–211.

Howson, C., & Urbach, P. (1993). *Scientific reasoning: The Bayesian approach*. Chicago: Open Court.

Johnson, D. M. (1972). *A systematic introduction to the psychology of thinking*. New York: Harper & Row.

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgement. In K. Holyoak & R. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–294). New York: Cambridge University Press.

Kao, S-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgement with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1363–1386.

Lo, Y., Sides, A., Rozelle, J., & Osherson, D. (2002). Evidential diversity and premise probability in young children's inductive judgement. *Cognitive Science*, *26*, 181–206.

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

Marascuilo, L., & Serlin, R. C. (1988). *Statistical methods for the social and behavioural sciences*. New York: Freeman.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, *10*(3), 517–532.

McDonald, J., Samuels, M., & Rispoli, J. (1996). A hypothesis-assessment model of categorical argument strength. *Cognition*, *59*, 199–217.

McKenzie, C. R., Ferreira, V. S., Mikkelsen, L. A., McDermott, K. J., & Skrable, R. P. (2001). Do conditional hypotheses target rare events? *Organisational Behaviour and Human Decision Processes*, *85*, 291–309.

McKenzie, C. R., & Mikkelsen, L. A. (2000). The psychological side of Hempel's paradox of confirmation. *Psychonomic Bulletin and Review*, *7*, 360–366.

McKenzie, C. R., & Mikkelsen, L. A. (2005). *A Bayesian view of covariation assessment*. Unpublished manuscript. University of California, San Diego.

McKenzie, C. R. M., & Mikkelson, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*, 33–61.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General*, *131*, 5–15.

Nickerson, R. S. (1996). Hempel's paradox and Wason's selection task: Logical and psychological puzzles of confirmation. *Thinking and Reasoning*, *2*, 1–31.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608–631.

Oaksford, M., Chater, N., & Larkin, J. (2000). Probabilities and polarity biases in conditional inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 883–899.

Oaksford, M., & Moussakowski, M. (2004). Negations and natural sampling in data selection: Ecological versus heuristic explanations of matching bias. *Memory & Cognition*, *32*, 570–581.

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

Siegal, M., & Surian, L. (2004). Conceptual development and conversational understanding. *Trends in Cognitive Sciences*, *8*, 534–538.

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

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd ed.). Oxford, UK: Blackwell.

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

Toppino, T. C., & Johnson, P. J. (1974). Interaction of positive and negative labels with category composition in attribute indentification concept performance. *Journal of Experimental Psychology*, *10*, 1035–1038.

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental Psychology*, *36*, 571–581.

Williams, P. B., & Carnine, D. W. (1981). Relationship between range of examples and of instructions and attention in concept attainment. *Journal of Educational Research*, *74*, 144–148.

APPENDIX A
Exemplars and properties used in Experiments 1a and 1b

Single evidence exemplar	*Positive or Negative evidence exemplar*	*Properties*	
Moose	Horse	*Biological*	
Kangaroo	Koala	has omit inside	has unti inside
Pig	Goat	has white bones	has yellow bones
Horse	Pig	is able to turn its head almost all the way around	can only turn its head around a little
Fish	Whale	needs a lot of calcium to stay healthy	only needs a little calcium to stay healthy
Snake	Alligator	has a 2-part heart	has a 3-part heart
Squirrel	Raccoon	has a paxtin stomach	has a fylate stomach
Monkey	Panda	has golgi blood	has filum blood
Deer	Fox	needs to eat a lot to stay alive	only needs to eat a little to stay alive
Wolf	Lion	has bunit ears	has liko ears
Rabbit	Beaver	has a round spleen	has an oblong spleen
Penguin	Walrus	has a round heart	has an oblong heart
Giraffe	Camel	grows new teeth every year	keeps the same teeth all year long
Cow	Sheep		
Turtle	Lizard	*Nongeneralisable*	
Bird	Cat	has a bug on it	does not have a bug on it
Frog	Mole	is 3 years old	is not 3 years old
Buffalo	Elephant	has a cut on its back	does not have a cut on its back
		has eaten food already	has not eaten food already
		has a brother	does not have a brother
		was born on a Thursday	was not born on a Thursday

APPENDIX B
Exemplars used in Experiment 2

Conclusion exemplar	*Subordinate negative*	*Basic negative*	*Superordinate negative*
Beagle (Dog)	Labrador	Cat	Robin
Grizzly Bear	Polar Bear	Moose	Dove
Swiss Cow	Angus Cow	Sheep	Alligator
Jack Rabbit	Cottontail Rabbit	Beaver	Lizard
Whitetail Deer	Starback Deer	Fox	Shark
Brown Squirrel	Grey Squirrel	Raccoon	Walrus

(*continued*)

APPENDIX B (*Continued*)

Conclusion exemplar	*Subordinate negative*	*Basic negative*	*Superordinate negative*
Mustang Horse	Norland Horse	Lion	Robin
Red Wolf	Grey Wolf	Zebra	Dove
Tamworth Pig	Duroc Pig	Mouse	Alligator
Tree Sparrow	House Sparrow	Dove	Cat
King Penguin	Rockhopper Penguin	Robin	Raccoon
Dart Frog	Tree Frog	Lizard	Mouse
Garter Snake	Rattlesnake	Alligator	Zebra
Painted Turtle	Snapping Turtle	Lizard	Beaver
Brown Trout	Rainbow Trout	Shark	Sheep
Humpback Whale	Pilot Whale	Walrus	Moose

APPENDIX C
Exemplars used in Experiment 3

	Negative premise			*Target conclusion*		
Positive premise	*Basic level*	*Superordinate*	*Kingdom*	*Same*	*Basic*	*Superordinate*
Cat	Deer	Fish	Plant	Cat	Dog	Bird
Cow	Rabbit	Fish	Plant	Cow	Horse	Bird
Squirrel	Elk	Fish	Plant	Squirrel	Mouse	Bird
Wolf	Monkey	Fish	Plant	Wolf	Fox	Bird
Pig	Camel	Fish	Plant	Pig	Sheep	Bird
Zebra	Raccoon	Fish	Plant	Zebra	Rhino	Bird
Elephant	Ferret	Fish	Plant	Elephant	Giraffe	Bird
Bull	Beaver	Fish	Plant	Bull	Moose	Bird
Tiger	Bear	Fish	Plant	Tiger	Lion	Bird