COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2012) 1–22

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2012.01257.x

How Young Children Learn From Examples: Descriptive and Inferential Problems

Charles W. Kalish, Sunae Kim, Andrew G. Young

Department of Educational Psychology, University of Wisconsin-Madison

Received 10 February 2011; received in revised form 2 January 2012; accepted 5 January 2012

Abstract

Three experiments with preschool- and young school-aged children (N = 75 and 53) explored the kinds of relations children detect in samples of instances (descriptive problem) and how they generalize those relations to new instances (inferential problem). Each experiment initially presented a perfect biconditional relation between two features (e.g., all and only frogs are blue). Additional examples undermined one of the component conditional relations (not all frogs are blue) but supported another (only frogs are blue). Preschool-aged children did not distinguish between supported and undermined relations. Older children did show the distinction, at least when the test instances were clearly drawn from the same population as the training instances. Results suggest that younger children's difficulties may stem from the demands of using imperfect correlations for predictions. Older children seemed sensitive to the inferential problem of using samples to make predictions about populations.

Keywords: Inductive inference; Cognitive development; Covariation; Belief revision; Statistical learning

Much of what we know about the world we have learned through encounters with examples. This learning process can be understood as having two parts or steps. First the learner notices some regularity in experience. One might observe that all of the birds encountered have been able to fly. In the second step the learner extends or generalizes that regularity beyond the examples encountered. One might infer that all birds fly. There is considerable debate over how to characterize the psychological mechanisms underlying this kind of inductive inference (Colunga & Smith, 2008; Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Xu, 2008), and how such inferences might develop (Sloutsky, 2003). Like many other researchers (Griffiths et al., 2010; Romberg & Saffran, 2010), we contend that learning from examples is best understood as a form of statistical inference. We note that

Correspondence should be sent to Charles Kalish, Educational Psychology, Rm 880b EdSciences, 1025 W. Johnson St., Madison, WI 53706. E-mail: cwkalish@wisc.edu

the two parts of learning from examples correspond to a fundamental distinction between descriptive and inferential statistics. Learning from examples presents two problems, a descriptive one and an inferential one. In characterizing performance, or development, on some learning task it is important to distinguish these problems. The current paper explores these issues in the context of changing expectations after encounters with novel examples. What happens after the learner encounters some flightless birds? We ask whether developmental differences in this kind of belief revision reflect changes in solutions to the descriptive problem or the inferential problem.

1. Descriptive problem: Conditional and biconditional relations

Most work on categorization and inference has explored the descriptive problem of learning from examples. How do people identify various kinds of relations present in the evidence they encounter? For example, noticing two clusters organized around prototypes is a way of describing the structure of a set of examples. That description might then be useful in making predictions about novel examples. Many research questions concern the relative difficulty of identifying various relations (e.g., are linearly separable categories easier to learn? see Murphy, 2002 for review) and developmental changes in children's abilities to learn different relations. In particular, work on ''statistical learning'' (see Romberg & Saffran, 2010) explores the kinds of statistical relations or patterns children notice in experience (e.g., transition probabilities and non-adjacent dependencies). Developmental differences in abilities to detect patterns may underlie differences in learning from examples.

The current study follows up on some suggestions that young children may find it more difficult to reason with simple conditional relations than biconditional relations. A biconditional is a symmetric relation of correlation between two variables. A biconditional can be understood as a conjunction of two ''simple'' conditionals. For example, ''All and only red fish live in warm water'' is equivalent to ''All red fish live in warm water'' and ''All warm-water fish are red.'' Research on conditional reasoning suggests that children tend to over-interpret statements of simple conditionals (''if red, then warm'') as statements of biconditionals (''if and only if red, then warm''; Barrouillet & Lecas, 2002). Additionally, preschool- and young school-aged children more readily learn biconditional relations in a set of examples than simple conditional relations (Kalish, 2010). For example, after encountering a set of examples in which all the red fish lived in warm water and all the green fish lived in cold, children reliably made predictions consistent with the biconditionals (e.g., presented a red fish, they predicted it would live in warm water). However, after encountering a set of examples in which only some conditional relations were reliable (all the red fish in warm water, but some green fish in warm and some in cold), children often failed to learn anything. That children are less likely to learn and use conditional relations than biconditional is consistent with the view that children encode general, ''gist,'' relations (e.g., red and warm ''go together''; Reyna & Brainerd, 1994), and that it is easier to learn associations (i.e., probabilistic biconditionals)

than specific conditional probabilities (Vadillo & Matute, 2007). Kloos (2007) notes that young children tend to assume that systems of relations are coherent, such that predictive relations holding between different variables are consistent with each other. Associative relations are symmetric (relation of A to B is the same as B to A), which may be an aspect of coherence. These considerations motivate the hypothesis that young children may fail to generalize a conditional relation to new examples because they have difficulty extracting or noticing that relation in familiar examples.

2. Inferential problem: Feature matching and evidence

The central focus of the current study is development of solutions to the inferential problem of learning from examples. The crux of the debate between different statistical approaches to cognition lies in their characterization of inferential abilities. Associative theories treat the inferential problem as one of matching, hence the designation ''similarity-based'' (Sloutsky, 2003). The relations observed in past examples (e.g., all birds have wings) will be extended to new examples based on the similarity of old and new exemplars. Empirical questions concern the nature of those similarity calculations (e.g., holistic or selective, based on perceptual or abstract features; Sloutsky & Lo, 1999). In contrast, theory-based and Bayesian accounts treat the relation between old and new exemplars as evidential (Gelman & Kalish, 2006; Griffiths et al., 2010; Xu & Tenenbaum, 2007a). A relation observed in old examples provides evidence regarding the existence of the relation in new examples. The evidential relation is formalized in the principles of inferential statistics. At least in large measure the debate between accounts turns on whether inference is a process of feature matching or a process of evidence evaluation. The current study addresses this question developmentally: Are there age-related differences in whether young children approach the inferential problem of learning from examples as one of similarity assessment or as one of evidence evaluation?

On Bayesian or theory-based views, a key piece of the inferential problem is understanding the process that generated the encountered examples. The characteristic feature of evidence evaluation is treating examples as samples. Which population the examples are informative about depends on how they were selected. Despite several recent demonstrations of infants' and young children's sensitivity to sampling (Xu & Denison, 2009; Xu & Tenenbaum, 2007a), there remains considerable debate about the need to incorporate such mechanisms in accounts of cognition and cognitive development. For example, connectionist models involve quite powerful mechanisms for discovering descriptive relations and matching old and new examples, but no representations of relations between samples and populations (Rogers & McClelland, 2004). We return to this question in the general discussion and consider the implications of sensitivity to evidential aspects of inferential problems for models of children's learning and belief revision. The remainder of this introduction introduces the task used in the empirical studies described in the paper and illustrates the specific descriptive and inferential problems posed.

3. Study overview

The experiments reported below present children with a consistent descriptive problem: After first encountering a biconditional relation (e.g., all and only red fish live in warm water) they see some discrepant instances that violate the biconditional but are consistent with some simple conditional relations (e.g., some red *and* some green fish live in warm water). The descriptive question concerns the kinds of relations that children will identify and use. If children are limited to representing biconditional relations, then they should either ignore the discrepant examples (and maintain their belief in a biconditional relation) or abandon all their prior beliefs (and respond randomly). In contrast, representing simple conditional relations allows a discriminative response to the discrepant examples. That is, the discrepant instances support some predictions (e.g., all red fish in warm water, all cold-water fish are green) but undermine others (e.g., all warm-water fish are red, all green fish in cold water). After experience with discrepant instances, will children maintain their beliefs in supported relations but abandon their beliefs in undermined relations? A necessary condition for doing so is having ''solved'' the descriptive problem of belief revision posed by the task.

While forming an accurate descriptive representation of the instances is necessary, it is not sufficient. There is still the inferential problem of using the representations to make further inferences. In the ''training phase'' of our task, children encounter a set of instances (in two blocks, early and late). In the ''testing phase'' children are asked to demonstrate what they have learned by generating predictions. The descriptive relation learned during the training phase provides a warrant for this prediction. If one has learned that ''all the red fish live in warm water'' during training, then upon encountering a red fish in the testing phase one ought to predict that it lives in warm water. However, this warrant depends on the relation between the training and test instances. That is, one ''ought'' to predict that a red fish will live in warm water if that fish is drawn from the same population as those encountered during the training phase. If the test fish is drawn from a different population, the relevance of the training instances is much less clear. The ''inferential'' problem of inductive inference is evaluating the relation between known cases (training) and unknown (test).

The current study addresses the inferential aspect of inductive inference by manipulating the evidential relation between the training and test instances. In the ''sample'' conditions the test instances are drawn from the training set: The children are making predictions just about the evidence. In the ''populations'' conditions the test items are novel instances from a larger population. The population conditions raise the possibility that the descriptive relations observed in the training instances might not hold for the testing instances. Our question is whether children will perform differently in the sample and population conditions. A child could fail to make the ''correct'' predictions about the training instances for two reasons. First, she might have difficulties representing descriptive relations in the training instances previously encountered. This difficulty would affect performance in both population and sample conditions. Alternatively, a child might be unable/unwilling to make the inferential leap to generalize from training instances to novel test instances. This difficulty would only affect performance in the sample conditions. The goal of manipulating the relation between

the training and test instances is to diagnose the source of difficulty in belief revision; is it descriptive or inferential.

In sum, the current study presents a series of belief revision problems. The first research question is whether young children are able to solve the descriptive problem of detecting relations by combining two different sets of instances (one instantiating a perfect biconditional and one not). Children will show their success by distinguishing between supported and undermined relations. Failure to distinguish supported and undermined relations is a null result that could be due to descriptive difficulties or due to the inferential structure of the problem. Thus, the second research question is whether young children approach the inferential problem posed by our belief revision tasks by considering evidential relations. Will children distinguish between supported and undermined inferences when the training and test items are drawn from the same population but not otherwise?

4. Experiment 1

4.1. Methods

4.1.1. Participants

Forty younger children (M = 4:8, Range = 4:1–5:2) and 40 older children (M = 7:11, Range = 7:2–8:9) participated. Forty-two participated in the Sample condition (21 younger, 21 older), and 38 in the Population condition. Children were recruited from daycares and afterschool programs serving a largely middle-class population in a mid-sized Midwestern city.

Design. Children learned about a set of instances composed of two binary features. Specifically, children heard about a set of fish, or a set of shells (with content counter-balanced). The fish varied in habitat (red or blue background) and color (green or white). The shells varied in shape (spiky or smooth) and pattern (plain or spotted). For ease of exposition we describe the structure of examples using the shell features. Children encountered instances in two training phases. In the Early phase, eight instances, four spiky-spotted shells and four smooth-plain shells presented a perfect biconditional relation between shape and pattern. In the Late phase, children encountered two more spiky-spotted shells, and importantly, six spiky-plain shells that partially undermined the established biconditional.[1]

After each training phase, children made a series of conditional predictions. There were four single conditional prediction questions, one for each feature. A child encountered a test instance known to have one feature, and then predicted the feature value on the other dimension. For example, ‘‘This shell is spiky. Do you think this shell is spotted or plain?’’ Participants selected an option and then indicated confidence on a three-point scale (‘‘just guessing,’’ ‘‘think maybe,’’ or ‘‘know for sure’’). No feedback was provided. Critically, instances encountered in the Late phase were inconsistent with the perfect biconditional observed in the Early phase but did support two conditional predictions, p(Spiky|Spotted) = 1, and p(Plain|Smooth) = 1 (see Fig. 1). These two conditional predictions are designated ‘‘Supported’’ by Late phase examples. The other two conditional predictions,

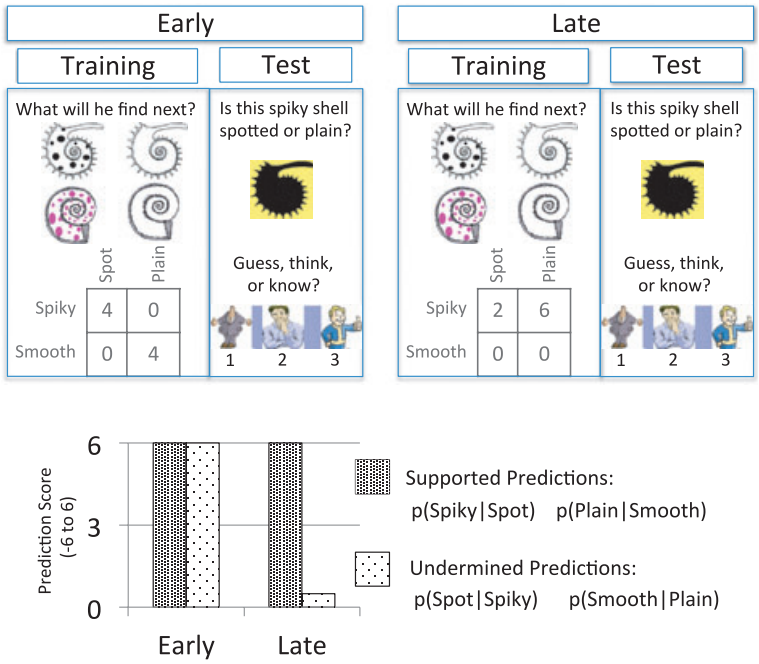

Fig. 1. Structure of Experiment 1 and expected scores for Supported and Undermined Predictions After Exposure to Early and Late Instances. High prediction scores mean reliable and confident prediction according to the biconditional relation in the Early instances.

p(Spotted|Spiky) and *p*(Smooth|Plain) were ''Undermined'' by the Late phase (both $p \approx .5$).

The inferential structure of the task involved the relation between training and test instances. In the Sample condition, the test instances were drawn from the set of training instances. That is, after each Learning phase, children were told that the shells or fish were collected into a box/tank. For example, children heard, ''Now Barney [the ''explorer'' in the story] is taking all the shells he found and putting them in this box.'' The conditional predictions involved instances drawn from the training set. ''Here is one of the shells from Barney's box. We know this one has spots, do you think it is spiky or smooth?'' In the Population condition, test instances were not part of the training set. The conditional predictions concerned new instances drawn from the larger population (e.g., Now Barney has found another shell on the beach. We know this one …). Children heard no information about the sampling procedure that generated the new test instances, just that they were not part of the training set. In all other respects, the Sample and Population conditions were identical.

4.1.2. Procedure

Instructions presented the task as a computer game in which explorers were learning about a newly discovered island. After a brief introduction that illustrated the dimensions and feature values of the set, children learned about a series of instances, one at a time. In

the first four trials of the Early training phase, the child watched as an explorer encountered four instances (randomly selected from either spiky-spotted or smooth-plain). The remaining Early phase trials began with pictures of the four possible instances. The child saw pictures of all four instance-types and then guessed which kind of instance the explorer would encounter next. After each instance appeared, the child moved it to the correct location for its kind. Locations were labeled (with words and a picture of the correct instance for the location). Following eight Early phase trials, the child made four single conditional predictions. Then, eight Late training phase trials were followed by another round of conditional predictions. All encountered instances remained visible throughout both training phases but were not visible during the prediction questions.

4.1.3. Scoring

Children's responses were coded as 1 when they were consistent with the biconditional relation presented during the Early training phase (spiky if and only if spotted), and as –1 when they were inconsistent with that relation. These codes were then scaled (multiplied) by the certainty rating (''guess'' = 1, ''think'' = 2, ''know'' = 3), resulting in predictions scores ranging from –3 to 3 (no zero). Data used in analyses were sums of the two supported predictions and sums of the two undermined predictions. These sums ranged from –6 to 6.

4.2. Results

Fig. 2 presents children's mean prediction scores for the Supported and Undermined conditional predictions. The key result is whether scores for Supported predictions were higher than scores for Undermined predictions after the Late training phase. Because the full design contains a large number of factors, and main results of interest would appear as three or

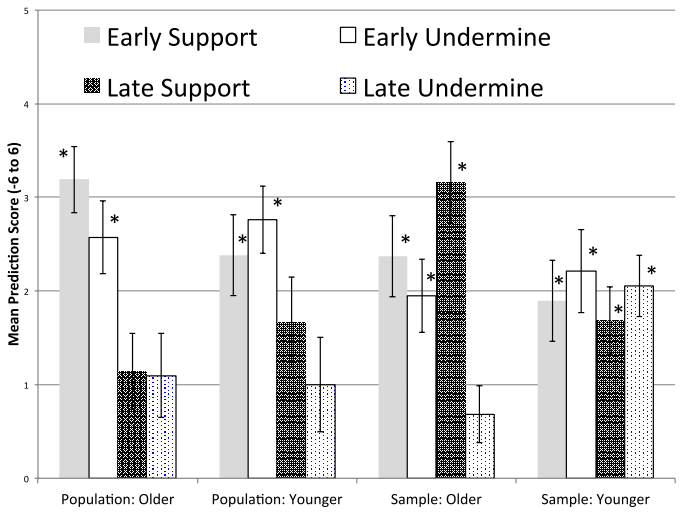

Fig. 2. Mean prediction scores, Experiment 1. Error bars indicate one standard error. *Mean score greater than chance, $p < .05$.

four-way interactions, data from the two conditions (Population and Sample) will be analyzed separately. Following this separate analysis we then compare across conditions.

4.2.1. Population condition

An ANOVA with Age as a between subjects-factor, and Support (Supported/Undermined) and Phase (Early/Late) as within-subjects variables revealed only an effect of Phase, $F(1, 40) = 4.4$, $p < .05$, ${}_{p}\eta^2 = .11$. Children gave lower prediction scores after seeing the Late phase instances. There was no effect of Support, $F(1, 40) = .2$, nor any interaction between Support and Phase, $F(1, 40) = .09$. There were no age differences. In general, children became less consistent in their predictions from the Early to Late phase, independent of the evidential support provided by the later instances.

Prediction scores combine both direction (consistent with the biconditional or not) and magnitude (confidence). Considering these components separately reveals similar patterns of results. Children gave fewer responses consistent with the biconditional in the late than the early phase for both supported (76–64%, $W(21) = 12$, $p < .05$, one-tailed) and undermined relations (75–58%, $W(24) = 114$, $p = .05$, one-tailed). In neither phase were rates of consistent predictions different for supported and undermined relations, and there were no age differences. We would expect that the number of children indicating they ''know for sure'' their predictions are correct would decrease from early to late phase for undermined but not supported relations, and the number of ''just guessing'' responses would increase. Children rarely indicated they were ''just guessing'' (average 13%), but often (average 59%) indicated they ''know for sure.'' These rates did not differ significantly by any of the factors (Age, Phase, or Support).

Children did learn the predictive relations supported by the Early phase instances. From Fig. 2 it is apparent that prediction scores were above chance in the Early phase. The same pattern appears in individual response patterns. Each child made four predictions. The chance of making all four consistent with the biconditional relation in the Early phase instances (spiky if and only if spotted) is .5^4 = .06. Twenty of 42 children showed this pattern. No child consistently predicted the biconditional pattern in the Late phase.

In summary, children in both age groups reliably used their experience in the Early phase to generalize a biconditional relation to novel instances. However, children abandoned this learning after seeing discrepant instances in the Late phase. Critically, there was no evidence of selective re-evaluation. Children stopped expecting Early phase relations that continued to be supported by Late phase experience as well as those that were undermined. Following discrepant evidence, children became very conservative, and seemed not to use their prior experience when predicting the properties of novel instances (chance-level performance). This effect could be a result of difficulty representing single conditional relations after Late phase experience. Alternatively, chance-level performance after the Late phase could be a result of concerns about the representativeness of the samples encountered in the training phases.

4.2.2. Sample condition

From Fig. 2 it is apparent that children responded quite differently to predictions about instances drawn from the training set (Sample condition) than to novel instances (Population).

ANOVA results from the Sample condition revealed an interaction between Age and Support, $F(1, 36) = 4.1$, $p = .05$, $\eta^2_p = .11$. Older children gave higher prediction scores for Supported than for Undermined predictions, $F(1, 36) = 5.3$, $p < .05$, $\eta^2_p = .18$. However, this difference was only significant following experience with discrepant instances in the Late phase, $F(1, 18) = 7.4$, $p < .01$. There was no reliable difference between Supported and Undermined predictions in the Early phase, $F(1, 18) = .88$. No other effects were significant in the ANOVA analysis. Critically, Younger children never distinguished between Supported and Undermined predictions.

The same general patterns appeared when direction and confidence of predictions are considered separately (rather than combined in prediction scores). Older children made more predictions consistent with the biconditional during the Late phase for Supported than Undermined relations (76% vs. 52%, $W(12) = 54$, $p < .05$, one-tailed). Younger children did not (72% vs. 75%). Children were generally confident, giving very few ''just guessing'' responses. However, Older children were less likely to ''know for sure'' about Undermined than Supported relations in the Late phase (63% vs. 39%, $W(10) = 39$, $p < .05$, one-tailed). Younger children did not show this confidence difference (50% vs. 62%). There were no differences on either measure for Supported and Undermined relations during the Early phase.

Children in both age groups learned and used the biconditional relation from Early phase experience (see Fig. 2). Younger children continued to base their predictions on the initial biconditional even after Late phase experience: Prediction scores were high, and greater than expected by chance, for both Supported and Undermined relations. In contrast, Older children reliably predicted Supported relations but responded at chance levels for Unsupported relations after Late experience. Fifteen participants (seven younger) made all four predictions consistent with the biconditional in the Early phase. Ten children (six younger) continued to show this pattern after the Late phase.

The most direct way to compare across conditions, for the result of interest, is to compare the differences in prediction scores for Supported and Undermined relations after Late experience. Did children distinguish supported from undermined relations? For Older children, the difference was relatively small in the Population condition, $M = .05$, but relatively large in the Sample condition, $M = 2.5$, $t(39) = 2.0$, $p = .05$. In contrast, Younger children's scores were similar for Supported and Undermined relations in both Population and Sample conditions, $M = .67$ and $M = -.4$, respectively, $t(39) = .81$.

4.3. Discussion

The results of Experiment 1 are fairly clear for the older, school-aged, children. These children learned a biconditional relation between two properties in an early sample of instances (e.g., spiky and only if spotted). Following experience with a later set of instances, older children adjusted their predictions to reflect the two simple conditional relations within the combined sample (e.g., spiky if spotted, not-spotted if not-spiky). However, given the same training sample, older children did not use the conditional relations to make predictions about new instances. This pattern of results suggests that school-aged children can

recognize simple conditional relations, but something about the task in the Population condition left them unwilling to generalize from the sample to a wider population. We suggest that the source of this reluctance lies in concerns about the representativeness of the early and later samples.

Older children did generalize to new instances following the early training phase. After seeing eight instances embodying a biconditional relation, children expected that relation would hold for new instances as well. All the spiky shells had been spotted, so the next spiky shell would be spotted as well. After encountering a later training set, which included some discrepant instances (e.g., smooth spotted shells), older children no longer generalized. From the children's perspective, the task presented two very different samples; the relation between the properties in the early set was very different than the relation in the late set. Which is the right one? Why are they different? With the representativeness of the samples in question, older children adopted a very conservative strategy and declined to generalize. Before the representativeness of the sample was called into question (in the early phase) children did reliably generalize to new instances.

The conclusions with respect to younger children are exclusively negative: Preschoolers failed to show that they represented simple conditional relations, as evidenced by not distinguishing supported from undermined predictions. One explanation for this failure is that the descriptive problem was too difficult for them. Representing biconditionals is rather easy, but simple conditionals are somewhat more difficult (see Kalish, 2010). However, extraneous task demands may have driven performance. This possibility is addressed in Experiment 2. Interestingly, there was some suggestion that younger children might also be sensitive to the inferential structure of the problem. When making predictions to new members of the population, the discrepant instances led to chance-level performance. However, when making predictions about the members of the observed sample, children tended to ignore the discrepant instances and continue to make predictions consistent with the biconditional. Perhaps young children's difficulty with the descriptive problem prevented them from responding to the inferential problem in the same way as older children.

It is important to note that Experiment 1 presented instances as pictures on a computer. In this design, children also encountered similar pictures that did not necessarily represent the observed instances (e.g., the options for guessing the next instance encountered, the illustrations of the correct ''boxes'' for encountered instances). It is possible that younger children were confused about which pictures represented instances and which did not. Children did learn from the early phase exposure, so they were not completely confused. Nonetheless, a procedure that more clearly established just which instances had been encountered might reveal better performance. Experiment 2 addressed this possibility by using physical toy instances. Another advantage of using real instances is that the relation between sample and population can be more clearly expressed. In Experiment 2 participants observed both early and late sets selected from the same population (i.e., a bag of toys). If older children were reluctant to generalize to a population in Experiment 1 because of concerns about sampling, Experiment 2 might assuage those concerns and support inferences about new instances. Would children generalize when it is clear that the new instances in the test phase were sampled in the same way, from the same population, as the instances in the training phase?

5. Experiment 2

5.1. Methods

5.1.1. Participants

Thirty-two younger children (Mean age = 5:1, Range 4:3–5:9) and 16 older children (Mean age = 7:11, Range 7:3–8:9) participated. Seventeen younger children participated in the Sample condition, 15 participated in the Population condition. Older children participated in the Population condition only. As Older children were predicted to distinguish between Supported and Undermined predictions in the Population condition, there was no value in including them in the (easier) Sample condition. Children were recruited from programs serving the same population as that of Experiment 1. No child participated in more than one experiment in this report.

5.1.2. Materials

Experiment 2 used three-dimensional toys in place of the computer-presented pictures from Experiment 1. Toys were soft rubber animals of about 2″ in length. The experimental set consisted of ten frogs and six dinosaurs. Each toy was of uniform color, either yellow or blue. Additional materials consisted of a large lidded ''source'' box and four ''evidence'' boxes, each displaying a picture of one of the toys.

5.1.3. Design

Experiment 2 followed the design of Experiment 1 very closely. Children saw eight Early phase instances exhibiting a perfect biconditional relation between color and species (four blue frogs and four yellow dinosaurs), and eight Late phase instances supporting some of the component conditional relations, but undermining others (six yellow frogs and two yellow dinosaurs). The conditional prediction questions and confidence responses took the same format as those in Experiment 1. We also assessed children's memory for the instances they encountered during the experiment. At the end of the experiment, children were asked, ''Which kind of toy did we see most?'' and ''Which kind of toy did we never have?''

5.1.4. Procedure

The procedure of Experiment 2 followed that of Experiment 1 with the exception that children did not make predictions about the objects they were to encounter. Each trial in both learning phases began with an experimenter pulling out a toy from the source box. The child then placed the toy into one of the four corresponding evidence boxes. As in Experiment 1, encountered instances were visible throughout both learning phases, but not during the prediction and memory questions.

5.2. Results and discussion

Fig. 3 shows children's mean prediction scores (calculated as in Experiment 1) for both Supported and Undermined relations. As Older children did not participate in the Sample condition, data from the two age-groups are analyzed separately.

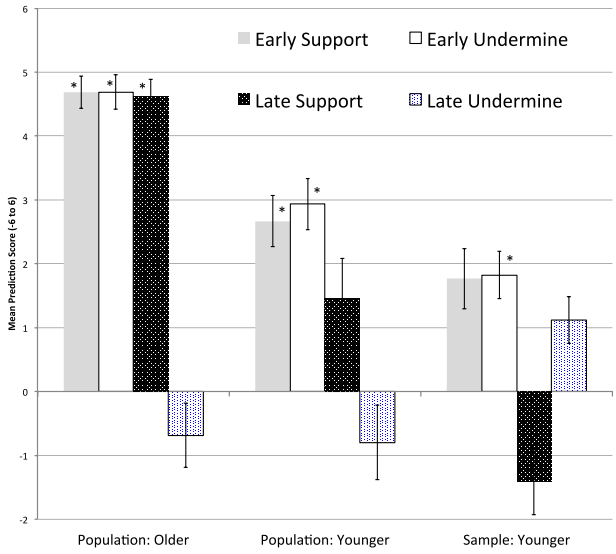

Fig. 3. Mean prediction scores, Experiment 2. Error bars indicate one standard error. *Mean score greater than chance, $p < .05$.

In contrast to the Population condition of Experiment 1, Older children did distinguish between Supported and Undermined relations. An ANOVA with Phase (Early/Late) and Support (Supported/Undermined) as within subject factors revealed a significant interaction, $F(1, 15) = 21.9$, $p < .001$, $\eta^2_p = .59$. Scores for Supported relations did not change from Early to Late phase, while scores for Undermined relations decreased, $F(1, 15) = 32.4$, $p < .001$. Similarly, Supported and Undermined scores did not differ after the Early phase, but scores for Undermined relations were significantly lower after the Late phase, $F(1, 15) = 29.3$, $p < .001$. Comparisons against chance-level performance support these conclusions. Older children's prediction scores were greater than expected by chance, except for Undermined relations in the Late phase (see Fig. 3). Older children did generalize the conditional relations in the full (Early + Late) sample to new instances. We suggest that the critical difference between Experiments 1 and 2 is that the sampling process in Experiment 2 was more clear, allowing children to recognize the relation between the sample of instances and the full population.

These patterns of results also appeared in direction and confidence judgments considered separately. Older children made more predictions consistent with the biconditional in the Late phase for Supported than for Undermined relations (93% vs. 44%, $W(11) = 66$, $p < .005$, one-tailed). They were also more likely to indicated they were ''just guessing'' for Undermined than Supported in the Late phase (47% vs. 13%, $W(10) = 43$, $p < .05$, one-tailed). Rates of claiming to ''know for sure'' did not differ significantly (Supported: 68%, Undermined: 44%) in the Late phase. However, children were significantly more likely to claim to know for sure about Undermined relations in the Early phase (82%, $W(9) = 40$, $p < .01$, one-tailed).

The physical stimuli and sampling process did not have the same effect for Younger children. An ANOVA with Condition (Sample/Population) as a between-subjects factor, and

Phase and Support as within, did reveal a main effect of Phase, with Early scores being higher than Late scores, $F(1, 30) = 9.7$, $p < .005$, $\eta^2_p = .24$. The effects of Phase and Support also interacted with Condition, $F(1, 30) = 10.1$, $p < .005$, $\eta^2_p = .26$. In the Population condition, there was no significant interaction between Phase and Support, just a main effect of higher prediction scores Early than Late, $F(1, 14) = 6.1$, $p < .05$, $\eta^2_p = .30$. This result is consistent with young children's performance in Experiment 1. Experience with the discrepant instances in the Late phase reduced all prediction scores in the Population condition. Similarly, young children's scores for both Supported and Undermined relations were greater than chance in the Early phase, but not in the Late phase (see Fig. 3). Thus, unlike Older children, the clarification of the sampling strategy in Experiment 2 did not lead Younger children to generalize the conditional relations present in the sample. There were no significant differences in rates of predictions consistent with the biconditional or in confidence ratings.

The hypothesis after Experiment 1 was that young children just had difficulty integrating across Early and Late training to identify the conditional relations in the sample (the descriptive problem). Would use of real objects help? Although children in the Sample condition showed the predicted interaction between Phase and Support, $F(1, 16) = 8.9$, $\eta^2_p = .36$, $p < .01$, the effect was in the opposite direction. Younger children gave significantly lower scores to Supported relations after Late phase experience, $F(1, 16) = 7.7$, $p < .05$. This pattern represents something like the Gambler's Fallacy: I have not seen any blue dinosaurs, so I will predict that the next blue thing will be a dinosaur (and the next dinosaur will be blue). Scores for Undermined relations did not differ between the Early and Late phases, $F(1, 16) = .49$. Low prediction scores in the Early phase contributed to this anomalous pattern. Children did not respond at greater than chance levels for Supported relations even in the Early phase (see Fig. 3), but they did do so for Unsupported. It is unclear why children did not respond reliably to Supported relations in the Early phase as (at that point) the evidential support for Supported and Unsupported inferences was identical. Though it is not exactly clear what did drive Younger children's predictions, the results are generally consistent across both experiments and across both conditions: Younger children did not use simple conditional relations to make predictions.

One explanation for young children's puzzling behavior may be that children in the Sample condition were simply not paying as much attention to the task. To answer the prediction questions correctly children simply needed to recall that they had seen instances of every kind but one (no blue dinosaurs). Children in the Population condition generally responded accurately to memory probes. Ninety-percent of children in both age groups identified blue dinosaurs as the kind of example they had never seen. Eighty-five percent of Older children correctly reported seeing the most of the yellow dinosaurs. Only 40% of Younger children were correct about the most frequent instance; responses were relatively evenly split among the three encountered types. Thus, children in the Population condition did encode the necessary information (though only older children seemed to use this information appropriately). Younger children in the Sample condition did not show robust memory: Only seven of sixteen correctly recalled which instance was never encountered, and nine recalled the most frequent instance (one child was not asked the memory questions).

Before considering the implications of Experiments 1 and 2 it is worth addressing one final alternative explanation for young children's chance-level performance following encounters with discrepant evidence. Perhaps the experience of two different samples is just confusing. Both older and younger children may experience the discrepancy between the Early and Late phase experience as problematic. Perhaps someone is being tricky. Perhaps something has changed. Older children are able to selectively respond to the discrepancy in the instances, and recognize that some reliable inferences are still possible. Young children might be responding to the inferential structure of the task, but in a very simple way; they are suspicious or confused about the samples, so revert to chance-level performance. One way to test this explanation of young children's performance is to present a task in which the discrepancy between the Early and Late phases is irrelevant: Both support the same conclusions. Would children continue to show chance-level performance even when the discrepant evidence was actually consistent with prior beliefs?

6. Experiment 3

Experiment 3 asked children to produce an outcome. For example, to get a dinosaur would they pick a yellow animal or a blue one? This task effectively calls for a judgment of relative conditional probability; is $p(\text{Dinosaur} \mid \text{Yellow}) > p(\text{Dinosaur} \mid \text{Blue})$. Although it would seem that this comparative judgment would be more difficult than evaluating the component conditionals, this may not be the case. The relative judgment is based on association, which may be a more automatic computation (Vadillo & Matute, 2007) or one that does not require selective attention (Kalish, 2010; Sloutsky & Fisher, 2008; Yao & Sloutsky, 2010). The important point, for the current study, is that the best selection does not change with discrepant evidence like that used in the prior experiments. As long as one cell in the contingency table is empty (all discrepant instances come from one of the off-diagonal cells), both Early and Late experience leads to the same behavior. For example, if one has never seen a blue dinosaur, then it will always be best to select a yellow animal in hopes of finding a dinosaur, even if yellow frogs greatly outnumber yellow dinosaurs. Thus, we predict that children will show the same performance in Early and Late testing phases: Encountering discrepant instances will not lead to chance-level performance in this task.

6.1. Methods

6.1.1. Participants

Nineteen younger children participated (M = 4:10, Range = 4:1–5:8). Children were recruited from programs serving the same population as that of Experiments 1 and 2.

5.1.2. Design and procedure

The structure of the task was nearly identical to that of Experiment 2. The only differences came in the prediction phases. Following each learning phase the experimenter pretended to sort all the instances into two boxes. All children responded to two questions

about the instances sorted by color and two by shape. For example, in the case of color the experimenter explained, ''I put all the yellow toys into this box and all the blue toys into this box.'' Each selection box had a picture of a corresponding feature (i.e., yellow/blue or frog/dinosaur). Children then selected one of the boxes to find a particular type of toy. For instance, while presenting two color boxes the experimenter asked, ''I want a frog. If I want a frog which box should I pick?'' Following their selection, children indicated how much better their box choice was than the alternative (''much better,'' ''a little better,'' ''does not matter''). Children made four judgments, one for each feature-value (frog, dinosaur, yellow, blue). Half the children received two color questions first and the other half received two shape questions first. At the end of the task, children received the same memory questions as those in Experiment 2.

6.1.3. Measure

Children's responses were converted to selection scores using the same procedures that generated prediction scores in Experiments 1 and 2. Selections consistent with the Early phase relation received positive scores. The ''how much better'' question determined magnitude: 1 for ''does not matter,'' 2 for ''a little better,'' and 3 for ''much better.'' Thus, scores for individual predictions ranged from –3 to 3. Following Experiments 1 and 2, two predictions were Supported by Late instances, and two were Undermined. For example, after the Late phase it is no longer certain that selecting the yellow box will produce a dinosaur (however, it remains infinitely more likely than getting a dinosaur from the blue box). Summing those two predictions yielded scores ranging from –6 to 6.

6.2. Results and discussion

Fig. 4 presents young children's mean selection scores for Experiment 3. An ANOVA with Phase and Support as within-subjects variables revealed no significant main effects, nor was the interaction significant, $F(1, 18) = 2.5$. In contrast to previous experiments, Late phase

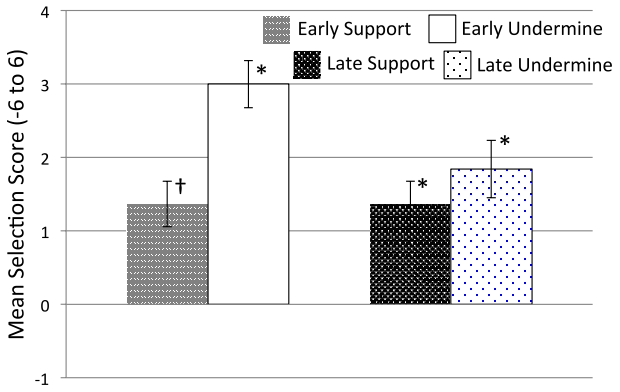

Fig. 4. Mean selection scores, Experiment 3. Error bars indicate one standard error. *Mean score greater than chance, $p < .05$. †$p < .05$, one-tailed test.

experience did not lead to a reduction in scores or to chance-level performance. Children gave selection scores that were significantly greater than chance in both Early and Late phases (see Fig. 4). As predicted, there were no differences between Supported and Undermined inferences. That some of the conditional predictions were undermined did not lead children to revert to chance level performance. Late phase experience left it impossible to predict the color of a frog (and children appreciated that fact in Experiments 1 and 2). Nonetheless, children reliably indicated that one was more likely to find a frog among one color of animal than the other. Although the Late phase presented some discrepant evidence, children realized that the set of instances encountered across both phases did allow some reliable inferences.

Just as in Studies 1 and 2, the selection scores combine direction (consistent with the biconditional observed in the Early phase) and confidence (how much better one selection is than the other). Considering these components separately did not reveal any significant Phase or Support differences. For example, children were equally likely to indicate selections consistent with the biconditional for Undermined relations in the Early and Late phases (71% and 82%, respectively). They were also equally likely to claim that their choice of selection was ''much better'' than the alternative for Undermined relations in both the Early and Late phases (39% and 42%, respectively).

Finally, children showed good recall. Fourteen of 16 correctly recalled which instance was never seen; 11/16 correctly recalled the most frequent. Three children were not asked memory questions because of experimenter error.

The results from Experiment 3 support the hypothesis that it is the descriptive task of identifying and using simple conditional relations that accounted for young children's chance-level performance in Experiments 1 and 2. The learning phases of Experiment 3 were identical to those of the Sample condition of Experiment 2. That children performed well in Experiment 3 but not in Experiment 2 suggests it is what they were asked to do with the examples that mattered. Experiment 3 required only noticing a general gist or overall association, and children performed well.

7. General discussion

The current study explored children's inductive inferences in the context of partially disconfirming evidence. After encountering a biconditional relation (all and only dinosaurs are blue) children encountered some partially disconfirming instances (all blue animals are dinosaurs, but not all dinosaurs are blue). Preschool-aged children generally failed to distinguish between those predictions that were supported by the discrepant instances and those that were undermined. Young school-aged children seemed to recognize the distinction but only used the past experience to make predictions when the relations between encountered and test instances were clear. Overall, the results suggest that young children had difficulty with the descriptive problem in the belief revision tasks used. Older children's performance reflects attention to the inferential problem of belief revision.

The current study supports previous findings that young children have difficulty using or representing simple conditional relations (Kalish, 2010). Young children find it easier to represent the overall gist (Reyna & Brainerd, 1994) or association between features, and they may have difficulty focusing their attention on the component relations that make up an association (Sloutsky & Fisher, 2008; Yao & Sloutsky, 2010). Of course, the current study presents only negative evidence: We failed to find evidence that young children made predictions based on simple conditional rather than biconditional relations. It remains possible that young children would show better performance under other conditions (e.g., with more training). However, the pattern of results support the claim that conditional relations are, at least, relatively difficult for young children. Young children did, however, reliably detect and use biconditional relations (early phases of Experiments 1 and 2), given even fewer examples than those available for simple conditional predictions (late phases). Moreover, younger children's chance-level performance cannot be attributed to the contrast between early and late training sets. In Experiment 3, younger children did make reliable judgments based on the overall associations between features in the full training set. It appears that the task of distinguishing evidential support for one conditional relation versus others is difficult for young children. Making such distinctions is part of the descriptive problem of belief revision.

It is important to note that the current studies do not identify the source of young children's difficulties in representing simple conditionals. One possibility is a representational limitation. For example, young children may lack the ability to focus on one ''direction'' of a relation. Vadillo and Matute (2007) suggest that the information processing demands of learning a conditional probability are greater than those involved in learning an association. An alternative view, though, is that children's performance reflects a bias or preference. Young children may tend to assume that relations are symmetric. They are able to represent conditionals, but they think conditional relations are rare. Presented with two binary features, children may assume the features are correlated. Commitment to this assumption (a strong symmetry prior) would lead children to have difficulty abandoning the belief in the face of disconfirming evidence. We tend to favor this second alternative, especially given evidence that infants are able to learn simple conditional relations (e.g., transition probabilities, see Romberg & Saffran, 2010). On this view it may be inappropriate to characterize young children as having ''difficulty'' with the task. Perhaps their performance represents an appropriate weighing of evidence with a (strong) prior. However, a strong prior on symmetric associations would not seem relevant in the sample conditions (where one is making predictions just about the observed examples). It may be that failure to appreciate the difference between sample and population inferences is the real source of young children's difficulties. That is, young children held to their prior expectation of biconditional relations even when such a prior expectation was less relevant.

Descriptive problems (representational, or strong symmetry prior) do not seem to account for school-aged children's performance. Rather the best explanation for older children's performance reflects the inferential structure of our tasks. When the relation between the training instances and test instances was clear (Sample condition of Experiment 1 and in Experiment 2), older children used the distribution of features in the training sample to

make predictions about a broader population (regardless of whether those distributions were biconditional or simple conditionals). However, older children failed to use the training examples to make predictions about new instances after encounters with discrepant instances in the Population condition of Experiment 1. We suggest that this condition posed a distinctive inferential problem. Children had encountered two very different samples (different distributions of features). There was no explanation for this difference. Was it random? Were the samples drawn from different populations? Given this uncertainty, children were unwilling to generalize to new instances.

In fact, children's reluctance was well founded. They were encountering a non-stationary process; the design of the experiment depended on samples shifting between early and late phases. This shift was hidden in Experiment 2 (the experimenters used trickery to make it look like instances were selected at random). Most interestingly, this uncertainty/trickery was irrelevant in the Sample conditions because children were asked to make predictions about members of the training sets. We are not claiming that children ''solved'' the inferential problem presented in Experiment 1. It is very difficult to give a normative account of just what one should predict given a non-stationary process like that used in the experiments, though ignoring the evidence (which children did) is as reasonable a response as any. Rather the central point is that young school-aged children were attentive to the inferential structure of the problem; they did not approach the task as simply a descriptive problem. Note we are not claiming that preschool-aged were insensitive to inferential structure. Their difficulties with the descriptive problem would have made any sensitivity difficult to detect.

That children treat the problem of learning from examples, of making inductions, as inferential is consistent with recent Bayesian approaches to cognition (Griffiths et al., 2010; Oaksford & Chater, 2007) and stands in contrast to most work in the field, especially in the developmental literature, which has focused almost exclusively on descriptive problems. For example, prototype and instance models of categorization concern the kinds of descriptive representations people form from encounters with instances (roughly, parametric or non-parametric, see Murphy, 2002). Similarity-based, associative, or statistical learning models explore the patterns people detect in experience (Xu, 2008). Associative accounts of generalization, of how patterns are extended to new cases, can be characterized as ''transductive'' rather than truly inductive. One implication of the current study is that such models are insufficient to account for children's learning and inference.

''Transductive inference'' is usually taken to describe a movement from instance to instance (Inhelder & PIaget, 1958), but it has been given a more precise formulation in machine learning by Vapnik (1998). In our terms, transduction is an inference strategy that ignores the distinction between samples and populations. That is, transductive inference does not distinguish between making a prediction about one of the N instances already encountered and a new ($N + 1$) instance. Transduction is a kind of simplifying assumption; it works well when the N sample is truly representative of the population that generated the $N + 1$ instance (see discussions of semi-supervised learning; Zhu & Goldberg, 2009). This simplification has been (implicitly) adopted by most psychological accounts of inference and categorization.

To illustrate transductive inference, and the transductive nature of psychological models, consider the following example. Imagine a learner has acquired a decision rule for assigning two class labels to a set of *N* objects. For example, there is a prototype of the Class 1 objects and a prototype of the Class 2 objects. What would the learner do when one of the *N* objects is re-presented without its class label? Clearly the best way to assign the label is to apply the learned decision rule; the learned rule is, in part, derived from this object. Matching the unlabeled object to a prototype is guaranteed to provide the most accurate prediction of class label achievable by the learner[2] because that is how the prototypes were formed. Psychological theories differ in their accounts of the decision rules people actually use (e.g., prototype or exemplar-based), and formal theories may differ in their accounts of which decision rule is really optimal. However, psychological theories tend not to differ in their accounts of what people do when encountering a novel unlabeled object. The process is the same: People apply the learned decision rule to the new object. Here, though, the justification is more complex. The new object played no role in determining the decision rule. Using the old rule for the new object is only a good idea if the new object is somehow related to the original *N* objects: if it is a member of the same population. Transductive inference is insensitive to the relation between old and new instances: It applies the old decision rule to both. Inductive inference is sensitive to the relation and can potentially modify the decision rule applied to new instances. Effectively, transductive inference treats the Population and Sample conditions of the experiments above as identical. The conditions differ only in terms of the relation between training and test instances.

One hypothesis is that transductive inference is a sufficient model of human inference. Indeed, research suggests that adults are quite good at computing the descriptive statistics of samples (learning decision rules), but quite poor at considering the relations between samples and populations (Fiedler, Brinkmann, Betsch, & Wild, 2000; Juslin, Winman, & Hansson, 2007). For example, people are very willing to make inferences from non-representative samples and do not seem to adjust for, or even notice, threats to sample validity.

If adults do not distinguish samples from populations, then it would be very reasonable to assume that young children do not either. Appreciating inferential statistics seems to be a relatively advanced aspect of reasoning. The debate between similarity-based and theory-based accounts of children's cognition can be understood as a disagreement about whether children represent broader evidential relationships, such as those between samples and populations, or whether they just reason from the particular set of encountered instances (Kalish & Lawson, 2007; Sloutsky, 2003; Sloutsky & Fisher, 2004). Of course, claims about poor inferential reasoning in adults have been challenged (Griffiths & Tenenbaum, 2005), and there is evidence that adults do consider relations between samples and populations when drawing inferences from evidence (Lawson & Kalish, 2009; McKenzie & Mikkelsen, 2007). The current study contributes to a growing body of literature suggesting that children also reason inductively; they show at least some sensitivity to inferential relations between samples and populations (Gweon, Tenenbaum, & Schulz, 2010; Kushnir, Xu, & Wellman, 2010; Xu & Tenenbaum, 2007b).

Summary

Learners face a number of challenges when generalizing relations observed in a set of examples. One class of problems is descriptive: The learner must identify the relations instantiated in the examples. In the current study descriptive demands involved moving from a perfect biconditional correlation to a set of partial, simple conditional relations. The results suggest that this descriptive demand is too great for preschool-aged children. This performance is consistent with young children's difficulties learning simple conditional relations in the absence of revision (Kalish, 2010). It remains for future studies to determine whether these difficulties reflect representational limitations or strong assumptions about biconditional relations. Whatever their sources, young school-aged children were able to handle the descriptive demands of the present belief revision task.

The second class of problems in belief revision is inferential: How should new evidence be brought to bear on prior beliefs? When this inferential problem was simplified by making sampling procedures more clear (i.e., Experiment 2) or eliminated by restricting inferences to the set of observed instances (i.e., the Sample conditions), school-aged children did make reliable predictions supported by the evidence. Certainly there are more and less sophisticated ways of dealing with sample characteristics, and it remains for future research to describe the details of children's abilities in this regard. The current study illustrates that school-aged children at least recognize inferential problems. They may depend on transductive inference but seem to appreciate some limits of this strategy.

The current study explored one kind of learning problem: changing a relatively simple belief in light of discrepant evidence. Preschool-aged children seemed unwilling or unable to abandon their prior belief in biconditionals: They did not describe the examples in terms of simple conditional relations. For school-aged children the new evidence also raised the complication of sample selection. Encountering two samples with quite different distributions of features (different descriptive statistics) seemed to introduce doubt about the population. That older children recognized this second complication suggests they understand belief revision as involving inductive inference. Recognizing the scope of the problem children see themselves facing will place researchers in a much better position to understand children's solutions.

Notes

1. The actual ''discrepant'' instances were counterbalanced across participants (e.g., half saw six spiky-plain shells, while half saw six smooth-spotted shells).
2. Given the learnable decision rules. This example imagines a learner limited to prototype representations.

Acknowledgments

This research was supported by a grant from the National Science Foundation (DLS/DRM) 0745423 to the first author. Thanks to Jordan Thevenow-Harrison and Rory Raasch for their help with data collection.

References

Barrouillet, P., & Lecas, J. F. (2002). Content and context effects in children's and adults' conditional reasoning. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *55*, 839.

Colunga, E., & Smith, L. B. (2008). Knowledge embedded in process: The self-organization of skilled noun learning. *Developmental Science*, *11*, 195–203. doi: 10.1111/j.1467-7687.2007.00665.x

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General*, *129*, 399.

Gelman, S. A., & Kalish, C. W. (2006). Conceptual development. In D. Kuhn, R. S. Siegler, W. Damon, & R. M. Lerner (Eds.), *Handbook of child psychology, Vol. 2. Cognition, perception, and language* (6th ed., pp. 687–733). Hoboken, NJ: Wiley.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*, 357–364. doi: 10.1016/j.tics.2010.05.004

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *107*, 9066–9071. doi: 10.1073/pnas.1003095107

Inhelder, B., & PIaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Juslin, P., Winman, A., & Hansson, P. (2007). The naive intuitive statistician: A naive sampling model of intuitive confidence intervals. *Psychological Review*, *114*, 678–703. doi: 10.1037/0033-295x.114.3.678

Kalish, C. W. (2010). How children use examples to make conditional predictions. *Cognition*, *116*, 1–14.

Kalish, C. W., & Lawson, C. A. (2007). Negative evidence and inductive generalization. *Thinking and Reasoning*, *13*, 394–425.

Kloos, H. (2007). Interlinking physical beliefs: Children's bias towards logical congruence. *Cognition*, *103*, 227–252. doi: 10.1016/j.cognition.2006.03.005

Kushnir, T., Xu, F., & Wellman, H. M. (2010). Young children use statistical sampling to infer the preferences of others. *Psychological Science*, *107*, 1084–1092.

Lawson, C. A., & Kalish, C. W. (2009). Sample selection and inductive generalization. *Memory & Cognition*, *22*, 651–670.

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

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

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. New York: Oxford University Press.

Reyna, V. F., & Brainerd, C. J. (1994). The origins of probability judgment: A review of data and theories. In G. Wright & P. Ayton (Eds.), *Subjective probability*. (pp. 239–272). Oxford, England: John Wiley & Sons.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA, MIT Press.

Romberg, A., & Saffran, J. R. (2010). Statistical learning and language acquisition. *WIRES: Cognitive Science*, *1*, 906–914. doi: 10.1002/wcs.78

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in Cognitive Sciences*, *7*, 246–251.

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

Sloutsky, V. M., & Fisher, A. V. (2008). Attentional learning and flexible induction: How mundane mechanisms give rise to smart behaviors. *Child Development*, *79*, 639–651. doi: 10.1111/j.1467-8624.2008.01148.x

Sloutsky, V. M., & Lo, Y.-F. (1999). How much does a shared name make things similar? Part 1. Linguistic labels and the development of similarity judgment. *Developmental Psychology*, *35*, 1478–1492. doi: 10.1037/0012-1649.35.6.1478

Vadillo, M. A., & Matute, H. (2007). Predictions and causal estimations are not supported by the same associative structure. [Empirical Study Quantitative Study]. *The Quarterly Journal of Experimental Psychology*, *60*, 433–447.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.

Xu, F. (2008). Rational statistical inference and cognitive development. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 3: Foundations and the future, pp. 199–215). New York: Oxford University Press.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*, 97–104.

Xu, F., & Tenenbaum, J. B. (2007a). Word learning as Bayesian inference. *Psychological Review*, *114*, 245–272.

Xu, F., & Tenenbaum, J. B. (2007b). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*, 288.

Yao, X., & Sloutsky, V. M. (2010). Selective attention and development of categorization: An eye tracking study. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the XXXII annual conference of the cognitive science society* (pp. 842–847). Mahwah, NJ: Erlbaum.

Zhu, X., & Goldberg, A. B. (2009). *Introduction to semi-supervised learning*. San Rafael, CA: Morgan & Claypool.