Rational Statistical Inference and Cognitive Development

Fei Xu

University of British Columbia

(In press, P. Carruthers, S. Laurence, and S. Stich (eds.), The innate mind: foundations and the future, vol. 3. Oxford University Press.)

Please address correspondence to Fei Xu, 2136 West Mall, Department of Psychology, University of British Columbia, Vancouver, B.C., Canada V6T 1Z4, or email to fei@psych.ubc.ca. Telephone: 604-822-5972. Fax: 604-822-6923.

All students of cognitive development agree that the central questions in development are 1) specifying the initial state of a human infant, 2) specifying the final state of development for a human adult, and 3) specifying how to get from the initial state to the final state. Then academic disputes ensue.

Cognitive developmental psychologists are roughly divided into two camps: those who are more or less nativists and those who are more or less empiricists. Some psychologists do not like these terms, and some alternatives are "those who believe in innate knowledge" and "those who believe in learning," or "those who believed in initial conceptual knowledge" and "those who believe in initial perceptual capabilities." This division is also correlated with whether a researcher believes in domain specificity or not: nativists tend to argue for domain-specific knowledge (even at the beginning of development) and domain-specific learning mechanisms; empiricists tend to argue for domain-general learning mechanisms that may result in domain-specific knowledge some years into development (for some representative explications of these views, see Carey & Spelke, 1994; Cosmides & Tooby, 1994; Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996; Hirschfeld & Gelman, 1994; Karmiloff-Smith, 1992; Gopnik & Meltzoff, 1996; Keil, 1989; Pinker, 1994; Smith, 2001; Spelke, 1994; among others).

Since Piaget was *the* developmental psychologist for much of the 20th century, his views were very much the mainstream and much of the literature on cognitive development in the last 20 years considered Piagetian conceptions of development as the starting point. Many researchers sympathetic to nativism have argued that Piaget was wrong in assuming that the infants were *tabula rasa* (or blank slates). Infants may indeed have object permanence very early in development and they may even possess systems of knowledge such as intuitive physics, intuitive psychology, and a language faculty that that embodies a universal grammar and a language acquisition device. Many empirical results have been reported to support this view, and some have suggested that much of later development is largely just enrichment (a la Plato or Chomsky). In contrast, researchers sympathetic to empiricism have argued that Piaget may still be fundamentally right about the initial cognitive state of the infants, and they offer alternative interpretations of the many nativists' demonstrations of early competence in infants. Furthermore, these researchers often emphasize the role of learning. They have reported many empirical studies to support the idea that infants and young children possess powerful learning mechanisms that allow them to gather statistical information from the environment and this is the basis for qualitative shifts in development. By providing demonstrations of learning mechanisms (be they associative, correlational, or whatever), these researchers argue that it is not necessary to posit innate knowledge. The high-level concepts and domain-specific knowledge we see later in development can emerge from perceptual primitives (a la Hume or Locke).

The dichotomy posed above between nativists and empiricists pits two things against each other: 1) how much innate knowledge is given and 2) how powerful the child's learning mechanisms have to be. The basic assumption is that if a lot of innate knowledge is given, then we need not worry too much about learning mechanisms or the role of input statistics; on the other hand, if very little innate knowledge is given, then we should focus on characterizing learning mechanisms and the role of input statistics from the environment.

There is no doubt that this dichotomy has generated much interesting theoretical and empirical work (for a clear explication and review, see Spelke & Newport, 1998), thus it has been useful in advancing the field of cognitive and language development. Nonetheless, many researchers have argued for a middle ground – after all, we all believe in some innate stuff (but we may disagree on whether we should call it "concepts" or "knowledge") and we all believe in learning (but we may disagree on whether learning is enrichment or whether learning can bring about fundamental changes in the child's conceptual system). The difficulty in taking the middle ground is that it is easily perceived as being wishy-washy. One reason is that researchers have not committed themselves to a set of learning mechanisms, or perhaps the types of learning mechanisms posited (e.g., correlational learning) seem relatively simple and perhaps insufficient for acquiring the representations and knowledge we see later in development. Without a strong commitment to what kinds of learning mechanisms are available to the child, it is difficult to spell out any details in answering the crucial question of how to get from the initial state to the final state of development.

In this paper, I advocate a view that is hopefully a substantive middle ground, one that commits us to a set of learning and inference mechanisms that may be critical for learning and development. I dub this view "rational constructivism." I appeal to mechanisms of statistical inference as a means to bridge the gap between discussions of innate knowledge and discussions of learning and conceptual change.

Why might this approach allow us to make progress towards a more comprehensive theory of cognitive development? One reason is that the fundamental tension between the nativist and the empiricist viewpoints is the lack of inductive inference mechanisms. Much of human learning in the real world is inductive learning, i.e., the learner makes generalizations or draws conclusions based on data, often times sparse or a relatively small amount of data. For example, a human child hypothesizes the meaning of a new word with just one or a few exposures (e.g., Quine, 1960; Bloom, 2000; Carey, 1978; Markman, 1989). A human child induces complex grammatical rules based on very little data, i.e., listening to the mature speakers around them for a couple of years (e.g., Gleitman, 1990; Pinker, 1989; Wexler & Cullicover, 1980). A human child learns the rules of physical support with only a few trials (e.g., Baillargeon, 2002; Wang & Baillargeon, 2005). A human child uses language to infer hidden properties of an object with just a few examples (e.g., Gelman, 2003). Although sometimes children do require many repetitions and a lot of data (e.g., learning the irregular past tense forms of English, memorizing the multiplication table), most of the time they are willing to make the inductive leap based on fairly limited amount of evidence. However, much of the literature on cognitive development lacks any commitment on what kinds of inductive inference mechanisms are available to the child and how these mechanisms may explain developmental changes. This gap in the literature may partially explain why the dialogue between nativists and empiricists has not gone very far over the years. The principal learning mechanism I appeal to is based on general principles of Bayesian inference, much studied in the philosophy of science (e.g., Howson & Urbach, 1989) and within psychology, in computational vision, reasoning, and language processing (e.g., Chater, Tenenbaum, & Yuille, 2006; Tenenbaum, Griffiths, & Kemp, 2006; Yuille & Kersten, 2006; see also Gigerenzer & Hoffrage, 1995).

What is Bayesian inference?

Bayesian inference is a formalism that allows a learner to combine prior knowledge (in the form of biases/constraints) with statistical information in the input in order to estimate how likely it is that a hypothesis (H) is true given the data (D) at hand. Here I put forth a simplified version of Bayes' rule to illustrate the conceptual point:

$$p\,(H|D) = \frac{p\,(H) \text{ x } p\,(D|H)}{p(D)}$$

(We can safely ignore p (D) because it is independent of H and it only serves to normalize the sum of all p(H|D) to be 1, that is, the hypotheses are mutually exclusive and exhaustive.)

Thus we are left with three components:

1) Priors, p(H): the probability of a hypothesis in the absence of any observed data. In order to assess p(H), the learner needs a hypothesis space, e.g., object categories as potential referents of count nouns. The computations include biases, constraints, and knowledge that a learner brings to a particular task or learning situation; they may be innately given or they may be learned (e.g., the shape bias in word learning);
2) Likelihood, p(D|H): the probability of the data given the hypothesis. This includes assumptions about how likely the data are observed if we make some educated guesses about the sampling condition (e.g., random sampling vs. non-random sampling). The statistical information in the input is critical in computing the likelihood.

3) Posterior, p(H|D): Combining priors and likelihood, we can derive posterior probabilities that give us a quantitative measure of how likely it is that a particular hypothesis is true given the observed data.

Why Bayesian inference? First, this is a well-studied mathematical formalism that gives us a principled way of combining prior knowledge and input statistics, and it has been particularly successful in computational vision, a branch of cognitive science and cognitive psychology. Second, it may provide a more satisfactory answer to the question "what are the learning mechanisms?" in cognitive development. *Prima facie* it seems a more promising candidate than standard associative learning (often implemented as connectionist networks) because a) it explicitly acknowledges the importance of prior knowledge (note this part may be innate or learned), b) it explicitly acknowledges the importance of input data (as reflected in the likelihood term), and c) it provides a principled way of combining the two. One of the problems with associative learning mechanisms is that it seems like a 'brute force' way of learning, contrary to what we know about animal or human learning. Bayesian inference, on the other hand, says that learners are able to employ "smart" learning mechanisms that allow them to make generalizations based on a fairly limited amount of data. Third, if we take the "child as scientist" metaphor seriously, the inference engine useful for scientific reasoning may be useful for studying development. Fourth, methodologically, by laying out the three components, we have a natural and explicit 'division of labor' that makes us to be more precise about our commitments as scientists.

To illustrate the basic idea of Bayesian inference, I borrow an example from Tenenbaum (1999). I will simplify the example somewhat for the purpose of explication.

Suppose you are told that a simple mathematical rule governs a set of numbers you will see that are between 1 and 100, e.g., odd numbers, even numbers, all numbers less than 25, all numbers between 37 and 68, powers of 2, all prime numbers less than 100, etc. You then observe some examples that are randomly drawn from a set of numbers that conforms to this simple rule. Let's say the first number you observe is 16, and you are asked to rate how likely one of the following rules may be the correct one: a) all even numbers, b) all odd numbers, c) all numbers between 2 and 60, d) all prime number less than 100, and e) powers of 2. It is clear two of the rules cannot be correct: b and d, since 16 is neither an odd number nor a prime number. As for the other three hypotheses, a, c, and e, you may feel reluctant to say which one of these is more or less likely to be the correct rule. After all, the one example you have seen, 16, is perfectly consistent with any one of the three rules. Now you observe a few more examples, 8, 32, and 4. Now the set of data you have to make your inference is much richer: 4, 8, 16, and 32. So which mathematical rule is most likely to be correct given a, c, and e? Again, the examples are consistent with all three rules, but I think most of us will say that e) "powers of 2" has become the most probable candidate. Why? What is the intuition behind the increase in confidence level (reflected in an increase in probability assignment) from seeing just one example to seeing a few examples?

What are the prior probabilities for the various hypotheses? Adults share intuitions about what counts as a likely hypothesis, e.g., all even numbers, all odd numbers, multiples of 3, numbers between 20 and 40, prime numbers, etc. In contrast, most of us would say that "all even numbers except 54" has a very low prior probability since it may be considered as "an unnatural rule." Similarly, "all powers of 2 except 4 and 64", "all even numbers plus 13", and many others also receive low prior probabilities for the same reason. That is, among a very large set of logical possibilities, some are considered a priori more likely than others. Some rules are psychologically more natural and plausible to us than others. This is not to say that we will never consider low probability hypotheses. Suppose we observe many examples, including 6, 8, 12, 14, 16, 18, 44, 56, 78, 92, and 13. We may have no choice but to conclude that the rule is most likely to be "all even numbers plus 13." In the face of a lot of data, we may begin to weigh the low prior probability hypothesis more and more. Importantly, we need a lot of data to convince ourselves that a low prior hypothesis is indeed the correct hypothesis.

How do we calculate the likelihood p(D|H) so we can combine it with the prior, p(H) to arrive at a posterior probability, p(H|D)? Our intuition says that although "all even numbers" is consistent with the set of observed examples (4, 8, 16, and 32), somehow "powers of 2" is a better candidate. It seems to us that it would be "a suspicious coincidence" that we would see these four specific examples if they were randomly chosen from the whole set of "all even numbers." Perhaps it is more likely that we would have seen something like "4, 8, 34, and 56" given the assumption of random sampling. On the other hand, there is nothing "suspicious" about seeing these four examples if they are randomly drawn from the set "all powers of 2." The mind is keen in detecting these "suspicious coincidences" (see many examples from visual perception, e.g., Knill & Richards, 1996) and this ability becomes part of the inference mechanism to allow us to make fairly accurate guesses about the structure of the world. In order to compute the likelihood, we take into account such "suspicious coincidences."

Now we can calculate the posterior probability p (H|D) from these two terms, p(H) and p(D|H). Since "all powers of 2" has a fairly high prior probability and a fairly high likelihood, the posterior probability is also high for this hypothesis. In contrast, even though "all even numbers" may have a fairly high prior probability, the likelihood for this hypothesis is lower due to the general principle of avoiding "suspicious coincidence." So the posterior probability will be lower than that of "all powers of 2." Importantly, the likelihood term is calculated based on the assumption that the examples the learner has observed are a random sample of the true hypothesis.

A case study in development: Learning words at different levels of a taxonomy

What is the evidence that language and cognitive development employs Bayesian inference mechanisms? With both adult and child learners, there is a growing body of research suggesting that in domains such as causal reasoning, property induction, sentence processing, word learning, and syntax acquisition, the behaviors of the learners can be best accounted for by assuming an implicit Bayesian inference mechanism (see Chater et al., 2006, Gopnik & Schulz, 2004, and Tenenbaum, et al. 2006 for reviews).

We have conducted two series of experiments with preschool children on how they acquire the meanings of words that refer to subordinate-level, basic level, and superordinate-level categories – a much-studied and much-debated topic in early word learning, and we have built computational models to account for the learning processes based on the principles of Bayesian inference (Tenenbaum & Xu, 2000; Xu & Tenenbaum, 2005, in press a, in press b).

Learning words at different levels of a hierarchy has traditionally been considered a challenge in the literature. Upon seeing a dog running by and somebody labeling it "A blicket!" the child learner faces a difficult induction problem. Does "blicket" refer to all and only dogs, all mammals, all German shepherds, this individual dog Max, all dogs plus all cats, all brown things, the front half of a dog, undetached dog parts, etc.? Psychologists have borrowed the philosopher Quine's (1960) famous under-determinacy problem as it applies to word learning. Despite this logical problem of induction, children learn words surprisingly rapidly and quite accurately. A 6-year-old child knows an average of about 6,000 words, and most of these are learned by simply observing the world and listening to mature speakers of the language around them (Bloom, 2000; Carey, 1982; Markman, 1989). How is such rapid learning possible?

Models for how children acquire the meanings of words traditionally fall into two classes. In Xu and Tenenbaum (in press a), we called one class of models "hypothesis elimination models" and the other class of models "associative learning models." Hypothesis elimination models treat the process of word learning as inferential in nature – the child is assumed to draw on a set of hypotheses about word meanings and to evaluate these hypotheses based on the input (e.g., Markman, 1989; Siskind, 1996). In contrast, associative learning models assume that the child keeps track of word-percept pairings and adjusts the strengths of these correlations based on repeated exposures (e.g., Colunga & Smith, 2005; Regier, 2003, 2005).

Proponents of the hypothesis elimination approach argue that prior constraints help the learner rule out many logically possible but psychologically implausible hypotheses. The whole object constraint, for example, rules out hypotheses such as undetached dog parts, and the taxonomic constraint rules out hypotheses such as all dogs plus all cats, or all brown things (Markman, 1989). After applying these two constraints, however, we are still left with the problem of choosing among subordinate, basic-level, and superordinate level categories, e.g., poodle, dog, and animal, since none of these three candidate word meanings violates the whole object or the taxonomic constraint. Thus an additional constraint is needed, namely the basic-level bias, which says that learners prefer to map words onto basic-level categories. By invoking the basic-level bias, the child is able to eliminate all the other hypotheses as candidate word meanings. However, children do learn words for other levels of the taxonomic hierarchy. We are now in need of further stipulations that would allow the child to learn words such as "poodle" or "animal". Psychologists have proposed special linguistic cues as one source of information to help the child out of this quandary. For example, parents may say, "See this? It is a poodle. A poodle is a kind of dog" (e.g., Waxman, 1990). It is not clear if such special linguistic cues are always available to children, but more generally, it is hard to imagine that for each word, the learner has to invoke special constraints in order to zoom in onto the correct meaning.

The associative learning models do not fare better, either. Existing models in this school tend not to be able to handle 'fast mapping' – the learner's ability to make a good guess about a word's meaning with one or a few positive examples (e.g., Markman & Wachtel, 1998; Carey & Bartlett, 1978) -- since the principal mechanism of learning is to keep track of word-percept pairings and adjust connection weights gradually. Once the word-percept pairings are established through many trials, these correlations can guide future generalizations of the new word (Colunga & Smith, 2005; Gasser & Smith, 1998; Regier, 1996, 2005; Smith, 2000).

I will argue for an alternative view that combines aspects of both approaches: the basic architecture is a form of rational hypothesis-driven inference, but the inferential logic is Bayesian and hence shows something of the graded statistical character of associative models (Xu & Tenenbaum, in press a). Confronted with a novel word, the learner constructs a hypothesis space of candidate word meanings and a prior probability distribution over that hypothesis space. Given one or more examples of objects labeled by the new word, the learner updates the prior to a posterior distribution of beliefs based on the likelihood of observing these examples under each candidate hypothesis.

In a word learning task, adults and 4-year-old children were given one or a few examples of novel words. In the one-example condition, each child received one example of a new word. The experimenter picked up an object in a pile, say a terrier, and labeled it a total of three times, “See? A fep!” In the three-example condition, each child received three examples of a new word. The experimenter labeled a total of three objects once each. The perceptual span of the three examples varied from trial to trial -- sometimes they were three slightly different terriers, or three different kinds of dog (e.g., a poodle, a Dalmatian and a terrier), or three different kinds of animal (e.g., a dog, a pelican, and a seal). Then both adults and children were asked to generalize the word to a set of new objects. We were interested in whether children would take into account both the number of examples (1 vs. 3) and the perceptual span of the examples (subordinate-level, basic-level, or superordinate-level).

Figure 1 shows the results from adults and children. We found that in the one-example condition, adults showed a generalization gradient dropping off at the basic-level and children showed a generalization gradient without much of a drop-off. In the three-example condition, both adults and children generalized to the most specific level of category that was consistent with the data. How would we account for these data in a Bayesian framework?

Insert Figure 1 about here.

To begin with, we constructed a hypothesis space based on adults' similarity judgments of the objects we used in the experiments. We used these ratings to construct a hierarchical tree that included various potential hypotheses for the meaning of a new word. Some candidates corresponded to subordinate, basic-level, and superordinate categories; some did not. To instantiate the idea of detecting 'suspicious coincidences,' we also computed the likelihood such that as the number of examples increases, more specific hypotheses (i.e., smaller ones) are preferred than larger hypotheses that are also consistent with the data. This fits with our intuition that if I were to teach a word such as "animal," it would be odd if I picked up three different dogs and labeled each of them with the word "animal." Similarly, if I were to teach a word such as "dog," it would be odd if I picked up three different terriers, labeled each, and ignored all the other kinds of dogs. That is, the learner makes the general assumption that she is getting a random sample from the true extension of the word. Figure 2 shows the model results given these assumptions (for more technical details, see Xu and Tenenbaum, in press a).

Insert Figure 2 about here.

These studies provide evidence that in a word learning task, children and adults make inferences according to the basic principles of Bayesian inference. Note that in our approach, no special constraints are needed to decide among a set of nested categories (subordinates, basic-level, and superordinates) and the phenomenon of fast-mapping is naturally accounted for in the model by assuming that the learner begins with a fairly small set of hypotheses and a powerful inference mechanism.

In a second set of studies (Xu & Tenenbaum, in press b), we replicated our previous results using novel objects and we presented a new model that takes into account a 'theory-of-mind' inference in the domain of word learning. One critical assumption in the Bayesian framework we presented here is the idea that the learner assumes a random sample. Here we manipulated sampling conditions to test this assumption more directly. In the teacher-driven condition, adults and 4-year-old children received three subordinate-level objects as the referents of a new word from the 'teacher'/experimenter. This is identical to the three-example subordinate condition of previous studies (Figure 3). In the learner-driven condition, however, the 'teacher' presented the learner with just one example of the new word. Then the learner was asked to pick two more examples, and critically the learner was told that if she got both examples right, she would get a sticker (a highly rewarding prize for preschoolers, and apparently for adults too!). In the latter condition, the learner eventually received three positive instances of the new word, but they have a different status from the three positive instances in the teacher-driven condition. The critical difference is that the 'teacher' knew the word but the learner did not. The learner was inclined to be conservative and pick out two more examples closest to the first one. The epistemic state of the learner was different from that of the ‘teacher,’ and we predicted that it was only in the teacher-driven condition that the learner would restrict their generalization to other subordinate examples, whereas in the learner-driven condition it would be the same if the learner had received just one example from the teacher. Figures 3 and 4 presented pictures of the novel objects and the results from the experiments as well as those from the model. I do not have the space to go over the model details here, but see Xu and Tenenbaum (in press b) for more discussion.

Insert Figures 3 and 4 about here.

Associate models often have trouble accounting for theory-of-mind inferences; the general tendency is for the proponents of this approach to try to explain away these inferences (as attentional tuning, for example, Smith, 2001). The classic hypothesis elimination approach takes these theory-of-mind inferences seriously, but it lacks a formal model to integrate these inferences with other constraints. Here we present the first step towards a Bayesian model that integrates prior constraints, input statistics, and theory-of-mind inferences.

Bayesian inference: A domain-general mechanism?

Is the inference mechanism we have investigated in word learning specific to language? It is unlikely given that various versions of the Bayesian formalism have been applied successfully in computational vision, causal reasoning, and language processing (Chater et al., 2006). Recently we have completed a property induction study to address the domain specificity issue in our task (Talbot, Denison, & Xu, 2007). We adopted the same experimental paradigm as the studies by Xu and Tenenbaum (in press a), except that instead of teaching the child a new word, we taught him/her a new property, e.g., this one has beta-cells inside. We used the same set of objects as before, and varied the number of examples (1 vs. 3) as well as the perceptual span of the examples (subordinate-level, basic-level, or superordinate-level). Results from 4-year-old children looked very similar to the results from the word learning studies. Children showed a generalization gradient with one example and sharpened their generalization function with three examples. With three examples, the children generalized to the most specific level that was consistent with the examples they have been shown. These findings suggest that the inference mechanism may be domain-general and further research is needed to test how broadly learners apply these principles.

Basic computational machinery in infants and children

Before we can use Bayesian inference to explain learning and development in various domains, one might ask whether there is any evidence that children possess any of the basic computational competences required by Bayesian statistics. Although a growing body of research suggests that infants, children, and adults can use powerful statistical learning mechanisms in language and visual learning, not much has been said or done with a particular formal inference engine in mind.

Two aspects of this mechanism have been investigated in our laboratory: random sampling and base rate information. A series of experiments with 8-month-old infants have shown that 1) they are able to understand (implicitly) that a sample that is drawn randomly from a population gives a good clue as to the composition of the whole population, and 2) if a population is skewed (base rate information), a random sample from this population will also be skewed (Xu & Garcia, under review).

These experiments employed the violation-of-expectancy looking time paradigm. Infants were seated in a highchair facing a puppet stage. They watched some events unfold, and at times they were shown outcomes that were either expected or unexpected given an adult interpretation of the events. The infant's looking times were recorded. The logic behind this method is that if infants had interpreted the events the way adults would, they should look longer at the unexpected outcome. Many studies in infant perception and cognition have successfully used this method in the last two decades (e.g., Baillargeon, 2002; Spelke et al., 1992).

In Experiment 1, we asked if 8-month-old infants could use the sample they were presented to make some guesses as to the composition of the overall population. After the infant was seated in the highchair, the experimenter brought out a small container with several red or white Ping-pong balls. The infant was handed a few Ping-pong balls one at a time, and were encouraged to hold them for a few seconds. This warm up phase was designed to give the infant some idea for what they might see later on the puppet stage. The experimenter returned behind the curtains and sat behind the puppet stage. Her upper body and her face were visible to the infant. After calibrating the infant's looking window for the observer, the experiment proper began with four familiarization trials. On each trial, a large opaque box was brought out, and its front panel was opened. On alternate trials, the infant saw either a box containing a large number of mostly red (with a few white ones mixed in) Ping-pong balls or a box containing a large number of mostly white (with a few red ones mixed in) Ping-pong balls. Across four familiarization trials, the total amount of redness and whiteness was equated for all infants. Infants were allowed to look at the content of the box on each trial until they turned away for 2 consecutive seconds. Eight test trials followed. On each trial, the experimenter brought out the same large opaque box and sat it down on the empty stage. She then brought out a small transparent empty container and placed it next to the large box. She picked up the large box and shook it a few times, and the content of the box made some noises. She then turned her head away from the box, closed her eyes, and reached into the box through a top slit. The slit was covered with white spandex and the experimenter was not able to see the content of the box through the slit (without pulling the spandex open deliberately). She pulled out one Ping-pong ball, say a red one, and placed it into the small transparent container. She shook the large box again, looked away, pulled out another Ping-pong ball through the top slit, and placed it in the small transparent container. This sequence of event was repeated a total of 5 times, after which the small transparent container had either 4 white and 1 red Ping-pong balls, or 4 red and 1 white Ping-pong balls (on alternate trials). The order in which the Ping-pong balls were pulled out was randomized. The experimenter then opened the front of the large box to reveal its content, either a box with mostly red Ping-pong balls or one with mostly white Ping-pong balls (a transparent barrier held the Ping-pong balls so they stayed inside the box but were visible to the infant). Half of the infants were shown the mostly red outcome and half the mostly white outcome on all test trials. The question was whether after seeing 4 white and 1 red Ping-pong balls being pulled out of the box, the infants would expect the box to contain mostly white Ping-pong balls. The basic assumption behind this expectation is that what the infant saw in the small transparent container was a random sample from the box. (A rating study with adults confirmed this intuition. After viewing these events, adults expected a box with mostly white Ping-pong balls if they had seen a sample of 4 white and 1 red, and vice versa if they had seen a sample of 4 red and 1 white Ping-pong balls). A total of 6 test trials were run. On alternate trials, either a sample of 4 white and 1 red or a sample of 4 red and 1 white were shown, and the outcome for a particular infant was either mostly red or mostly white for all test trials. Looking times for the outcomes were recorded. We found that infants looked longer at the unexpected outcome than the expected outcome, that is, the one that did not match the sample they had seen. Experiment 2 replicated this finding with a different ratio in the sample, 6:1 (6 white and 1 red, or 6 red and 1 white). These results suggest infants assumed that the sample they saw was a random sample from the population, therefore they could use the sample to make educated guesses about the composition of the overall population. In Experiment 3, we tested the reverse, i.e., whether 8-month-old infants could use simple base rate information to predict the composition of the sample. The experimental procedure was very similar to that of Experiments 1 and 2, except that at the beginning of each test trial, the front panel of the big box was opened to show its content, and the infants were given 5 seconds to look at it. Again, half of the infants saw the mostly red contents for all test trials and half the mostly white contents for all test trials. The front panel was then closed, and a sample was drawn from the box as before. On alternate test trials, a sample of 4:1 or 1:4 was drawn, and looking times were recorded after all 5 Ping-pong balls had been placed into the small transparent container. If infants could use base rate information to make predictions about the sample, they should look longer at the unexpected outcome of 4 white and 1 red than the expected outcome of 4 red and 1 white if they had been shown a box with mostly red Ping-pong balls, and vice versa if they had been shown a box with mostly white Ping-pong balls. The results were as predicted: the infants remembered the overall content of the big box and they looked longer when a low-probability sample was drawn from it. Experiment 4 replicated this finding with a different ratio, 6:1. Again, infants were able to use the base rate information and looked longer at the unexpected outcome that did not match the contents of the box. Several methodological cautions were taken to ensure that results reflected an (implicit) understanding of random sampling and base rate information: Since no habituation was used, it was not possible to argue that the infants had learned the correct answer from habituation; since the same amount of redness and whiteness was presented during the familiarization trials, it was also not possible to argue that the infants had been more habituated to one type of outcome than the other. How do we know that the infants in fact made a connection between the sample and the population? Or put it slightly differently, an alternative interpretation might be that the infants noticed the ratio of the sample (or the population, as in the base rate experiments), and whenever the ratio changes, it elicited longer looking times. In two control experiments, we showed that if the sample of Ping-pong balls came from the experimenter's pocket (and not the big box) and were placed in the small container, the infants did not look longer at the unexpected outcome. We suggest that the infants had reasoned about the sample and its relation to the population, and it was not just a change in ratio between red and white Ping-pong balls that elicited the longer looking times on the test trials of Experiments 1 through 4.

The results of these 6 experiments suggest that 8-month-old infants may already have an implicit understanding of the basic assumptions of Bayesian inference. Young infants were able to relate samples to populations and vice versa, making statistical inferences that seem to obey the basic laws of probability. Many follow-up studies are underway to address issues such as how fine-grained these computations are, e.g., are they heuristics or probabilistic forms of reasoning?

Similar experiments have also been conducted in our laboratory with preschoolers (Denison, Garcia, Konopczynski, & Xu, 2006; Denison & Xu, under review). A different procedure was employed but the basic design was similar. Four-year-old children were asked to play a game with a puppet and help the puppet answer some questions. In the first experiment, the children were shown pairs of boxes with a different mix of colored objects. For example, one pair of boxes contained yellow and blue dog bones. One box contained mostly yellow dog bones, and the other contained mostly blue dog bones. Then behind an occluder, the experimenter reached into one box and drew out a sample of dog bones. On some trials the sample consisted of 5 yellow and 1 blue dog bone; on other trials the sample consisted of 1 yellow and 5 blue dog bones. Then the child was asked to help the puppet decide which box the sample came from. In two experiments, the preschoolers chose the correct box 61% and 74%, respectively. Their performance was significantly better than chance (50%) in both experiments. Then we asked the converse question by showing the child the content of the box at the beginning of each trial, then asking the child to choose between two samples. The child was asked to decide which of the two samples came from the box and which came from the puppet's house. In two experiments, the preschoolers chose correctly 69% and 80%, respectively. Their performance was significantly better than chance (50%).

We also showed adults video clips that we presented to infants and children, and asked them to rate the outcomes as expected or unexpected on a 7 point scale. Adults had very clear intuitions about which outcome was unexpected and they behaved similarly to infants and preschool children.

Two important questions remained unanswered by these studies. First, are learners sensitive to sampling conditions? We assumed that the infants, children, and adults all took the sampling procedure as a random draw from the population, but we do not have any direct evidence that a different sampling procedure may produce different results. On-going studies try to address this issue by comparing a random sampling condition with one where the experimenter looked into the box and drew out the samples deliberately. As we have seen earlier, we do have some evidence that in the context of word learning, preschoolers are sensitive to sampling conditions and it has consequences for how far they are willing to generalize a new word. Second, what is the underlying computation in these studies? Is it something approximating probabilities or is it just a heuristic? On-going studies try to address this question by asking infants and preschoolers to make more fine-grained judgments with different ratios of Ping-pong balls or dog bones.

To recapitulate, we have reported several series of experiments suggesting that some of the most basic components of Bayesian reasoning might be present in infants and young children, as well as adults. Much work is needed to further specify the nature of these inference mechanisms.

Conclusions

In this paper, I have tried to advocate a view that is hopefully a substantive middle ground between the extreme versions of either nativism or empiricism – a view I dubbed "rational constructivism." This is a view that commits us to some innate (or acquired) constraints and a set of powerful learning and inference mechanisms that may be critical for development. I have appealed to mechanisms of statistical inference as a means to bridge the gap between discussions of innate knowledge and discussions of learning and conceptual change. In particular, I have adopted the general framework of Bayesian inference and presented some recent research providing empirical evidence for the psychological reality of these inference mechanisms.

Many questions remain open since this is the beginning of a new research program. For example, how does the learner construct the hypothesis space? Are people really Bayesian given much of the reasoning literature from the last few decades? Is the inference mechanism really domain-general? Could this learning and inference mechanism bring about conceptual change? I will try to give some tentative answers to these questions in turn.

How the learner constructs the hypothesis space in each learning situation is an extremely important question. I think one source for generating hypotheses in the case of learning words is the part of the learner's conceptual structure that is concerned with categories and kinds (Markman, 1989; Xu, 2005). If language learning is largely a mapping problem, then the inference mechanism discussed here provides a principled way of choosing among a set of concepts. Where do representations of categories and kinds come from? Some research suggests that these are acquired during the first year of life (e.g., Xu, 2002; Xu & Carey, 1996; Xu, Cote, & Baker, 2005), although some of our core concepts are perhaps innately given, e.g., the concept of an object (Spelke, 1990; Spelke et al., 1992). Although I emphasize learning here, the 'rational constructivist' view does not eliminate the need for innate concepts. The infant must start with a set of perceptual and conceptual primitives, and ways of generating new hypotheses.

Are people Bayesian? Although much of the reasoning literature suggests 'no' (see Kahneman, Slovic, & Tversky, 1982), many have argued in recent years that people are much more Bayesian than this literature suggests. For example, Gigerenzer, Chater, Cosmides and colleagues have provided many demonstrations that people can reason rationally when they are presented with tasks and formats that are more ecologically valid, and many of the findings from the heuristics and biases literature have been reinterpreted in terms of a 'rational analysis' (e.g., Cosmides & Tooby, 1996; Gigenrenzer & Hoffrage, 1995; Oaksford & Chater, 1996; among others). Furthermore, recent computational models of visual perception, causal reasoning, and inductive inference have shown that people's behaviors are best captured in a Bayesian framework (see special issue of Trends in Cognitive Sciences, 2006). One (perhaps obvious) point to make is that just like other computational mechanisms that have been discovered in the last few decades, the Bayesian inference mechanism is employed implicitly without conscious awareness. An analogy to language may make this point even more transparent: although most of us are fluent speakers of English, the underlying computations we carry out in order to understand or produce language are entirely opaque to us. Indeed, it has taken linguists and psycholinguists many years of research to specify these underlying mechanisms. Similarly, we reason and make decisions everyday but the underlying computational processes are just as opaque to us as the mechanisms of motion detection, walking, or language use. It is perhaps unsurprising that research has uncovered sophisticated mechanisms for reasoning as it has in the case of language production and language acquisition.

Is the Bayesian inference mechanism domain-general, and if yes, in what sense? I have suggested throughout this chapter that this is not a mechanism specific to word learning, or language, or causal reasoning. However, I am not claiming that the same token of the Bayesian inference mechanism is used again and again in various domains. Rather Bayesian inference is a type of learning mechanism that can be instantiated many times over in the human brain/mind (I thank Peter Carruthers for raising this point).

Lastly, can these inference mechanisms bring about conceptual change? Perhaps it is clear how learning proceeds within this framework. As for conceptual change, it is an open question. Some have suggested that applying Bayesian learning algorithms to Bayes nets (talk about terminological confusion!) may provide a tool for conceptual change – as learning proceeds, new variables can be postulated and integrated into an existing network (see Gopnik & Schulz, 2004).

I hope the reader is now convinced that a substantive middle ground is possible – one does not have to commit to extreme versions of nativism or empiricism in the study of cognitive and language development. Furthermore, my collaborators and I have suggested that infants and children already have a powerful set of learning and inference mechanisms that are Bayesian in character. This is the beginning of a new research program, and I hope it will be a fruitful and productive one for years to come.

References

Baillargeon, R. (2002) The acquisition of physical knowledge in infancy: a summary in eight lessons. In U. Goswami (ed.), *Blackwell Handbook of Childhood Cognitive Development (pp. 47-83).* Blackwell Publishing.

Bloom, P. (2000) *How children learn the meanings of words*. Cambridge, MA: MIT Press.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, and G.A. Miller (eds.), *Linguistic theory and psychological reality (pp 264-293).* Cambridge, MA: MIT Press.

Carey, S. & Bartlett, E. (1978) Acquiring a single new word. *Papers and Reports on Child Language Development, 15,* 17-29.

Carey, S., & Spelke, E. S. (1994). Domain-specific knowledge and conceptual change. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture*, pp. 169-200. Cambridge, UK: Cambridge University Press.

Chater, N., Tenenbaum, J.B. & Yuille, A. (2006) Probabilistic models of cognition: conceptual foundations. *Trends in Cognitive Sciences, 10,* 287-291.

Colunga, E. & Smith, L.B. (2005). From the lexicon to expectations about kinds: the role of associative learning. *Psychological Review, 112,* 347-382.

Cosmides, L. & Tooby, J. (1996) Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1-73.

Denison, S., Konopczynski, K., Garcia, V., & Xu, F. (2006) Probabilistic reasoning in preschoolers: random sampling and base rate. In R. Sun and N. Miyake (eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1216-1221). Mahwah, NJ: Erlbaum.

Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D. & Plunkett, K. (1996) *Rethinking innateness: a connectionist perspective on development.* Cambridge, MA: MIT Press.

Gasser, M. & Smith, L.B. (1998). Learning nouns and adjectives: A connectionist approach. *Language and Cognitive Processes, 13,* 269-306.

Gelman, S.A. (2003) *The essential child.* Oxford University Press.

Gigerenzer, G. & Hoffrage, U. (1995) How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review, 102,* 684-704.

Gleitman, L.R. (1990). The structural sources of verb meanings. *Language Acquisition, 1,* 3-55.

Gopnik, A. & Meltzoff, A. (1996) *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Gopnik, A. & Schulz, L.E. (2004) Mechanisms of theory-formation in young children. *Trends in Cognitive Sciences, 8,* 371-377.

Hirschfeld, L. & Gelman, S.A. (1994), *Mapping the mind: Domain specificity in cognition and culture*. Cambridge, UK: Cambridge University Press.

Howson, C. & Urbach, P. (1989) *Scientific reasoning: the Bayesian approach.* Chicago, IL: Open Court Publishing.

Kahneman, D., Slovic, P. & Tversky, A. (1982) *Judgment under uncertainty: heuristics and biases*. Cambridge University Press.

Karmiloff-Smith, A. (1992) *Beyond modularity*. Cambridge, MA: MIT Press.

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

Knill, D.C. & Richards, W. (1996) *Perception as Bayesian Inference.* Cambridge University Press.

Markman, E.M. (1989) *Naming and categorization in children.* MIT Press.

Markman, E.M. & Wachtel, G.F. (1988) Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology, 20,* 121-157.

Oaksford, M. & Chater, N. (1998) *Rational models of cognition.* Oxford University Press.

Pinker, S. (1989) *Learnability and cognition: The acquisition of argument structure.* Cambridge, MA: MIT Press.

Pinker, S. (1994) *The language instinct.* William Morrow.

Quine, W.V.O. (1960) *Word and object.* Cambridge, MA: MIT Press.

Regier, T. (2003). Emergent constraints on word-learning: A computational review. Trends in Cognitive Science, 7, 263-268.

Regier, T. (2005). The emergence of words: attentional learning in form and meaning. Cognitive Science, 29, 819-866.

Siskind, J.M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61,* 39-91.

Smith, L. B. (2000) Avoiding associations when it’s behaviorism you really hate. In R. M. Golinkoff and K. Hirsh-Pasek (eds.), *Becoming a word learner: A debate on lexical acquisition.* Oxford University Press.

Spelke, E.S. (1990) Principles of object perception. *Cognitive Science, 14,* 29-56.

Spelke, E.S. (1994) Initial knowledge: six suggestions. *Cognition, 50,* 431-445.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99*, 605-632.

Spelke, E. S., & Newport, E. (1998). Nativism, empiricism, and the development of knowledge. In R. Lerner (Ed.), *Handbook of child psychology, 5th ed., Vol. 1: Theoretical models of human development*. NY: Wiley.

Talbot, A., Denison, S. & Xu, F. (2007) Camshafts and chlorophyll: statistical information and category-based induction in preschoolers. Poster presented at the Society for Research in Child Development Biennial Conference. Boston, MA.

Tenenbaum, J.B. (1999). Bayesian modeling of human concept learning. Advances in Neural Information Processing Systems 11. Kearns, M., Solla, S., and Cohn, D. (eds). Cambridge, MIT Press, 1999, 59-68.

Tenenbaum, J.B., Griffiths, T.L., & Kemp, C. (2006) Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10,* 309-318.

Tenenbaum, J.B. & Xu, F. (2000). Word learning as Bayesian inference. In L. Gleitman and A. Joshi (eds.), Proceedings of the 22nd Annual Conference of the Cognitive Science Society (pp. 517-522). Hillsdale, NJ: Erlbaum.

Wang, S. & Baillargeon, R. (2005) Inducing infants to detect a physical violation in a single trial. *Psychological Science, 16,* 542-549.

Waxman, S.R. (1990) Linguistic biases and the establishment of conceptual hierarchies: Evidence from preschool children. *Cognitive Development, 5,* 123-150.

Wexler, K. & Cullicover, P. (1980) *Formal principles of language acquisition.* Cambridge, MA: MIT Press.

Xu, F. (2002) The role of language in acquiring object kind concepts in infancy. *Cognition, 85,* 223-250.

Xu, F. (2005) Categories, kinds, and object individuation in infancy. In L. Gershkoff-Stowe and D. Rakison (Eds.), *Building object categories in developmental time: Papers from the 32nd Carnegie Symposium on Cognition (pp. 63-89).* New Jersey: Lawrence Erlbaum.

Xu, F. & Carey, S. (1996) Infants' Metaphysics: The case of numerical identity. *Cognitive Psychology, 30,* 111-153.

Xu, F., Cote, M., & Baker, A. (2005) Labeling guides object individuation in 12-month-old infants. *Psychological Science, 16,* 372-377.

Xu, F. & Garcia, V. (2007) Intuitive statistics by 8-month-old infants. Manuscript under review.

Xu, F. & Tenenbaum, J.B. (2005) .Word learning as Bayesian inference: evidence from preschoolers. In B.G. Bara, L. Barsalou, and M. Bucciarelli (eds.), Proceedings of the 27th Annual Conference of the Cognitive Science Society (pp. 2381-2386). Mahwah, NJ: Erlbaum.

Xu, F. & Tenenbaum, J.B. (in press a). Word learning as Bayesian inference. Psychological Review.

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

Yuille, A. & Kersten, D. (2006) Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10,* 301-308.

Acknowledgment

I thank Susan Birch, Paul Bloom, Susan Carey, Geoff Hall, Elizabeth Spelke, Janet Werker, and especially Josh Tenenbaum for many helpful discussions. I also thank Peter Carruthers for comments on an earlier version of the manuscript. Thanks to members of the UBC Infant Cognition Laboratory for their assistance in data collection. This research was supported by grants from the Natural Science and Engineering Research Council of Canada (NSERC) and the Social Science and Humanities Research Council of Canada (SSHRC). Address correspondence to Fei Xu at Department of Psychology, 2136 West Mall, University of British Columbia, Vancouver, B.C., V6T 1Z4, Canada. Phone: (604) 822-5972. Fax: (604) 822-6923. Email: fei@psych.ubc.ca.

Figure caption.

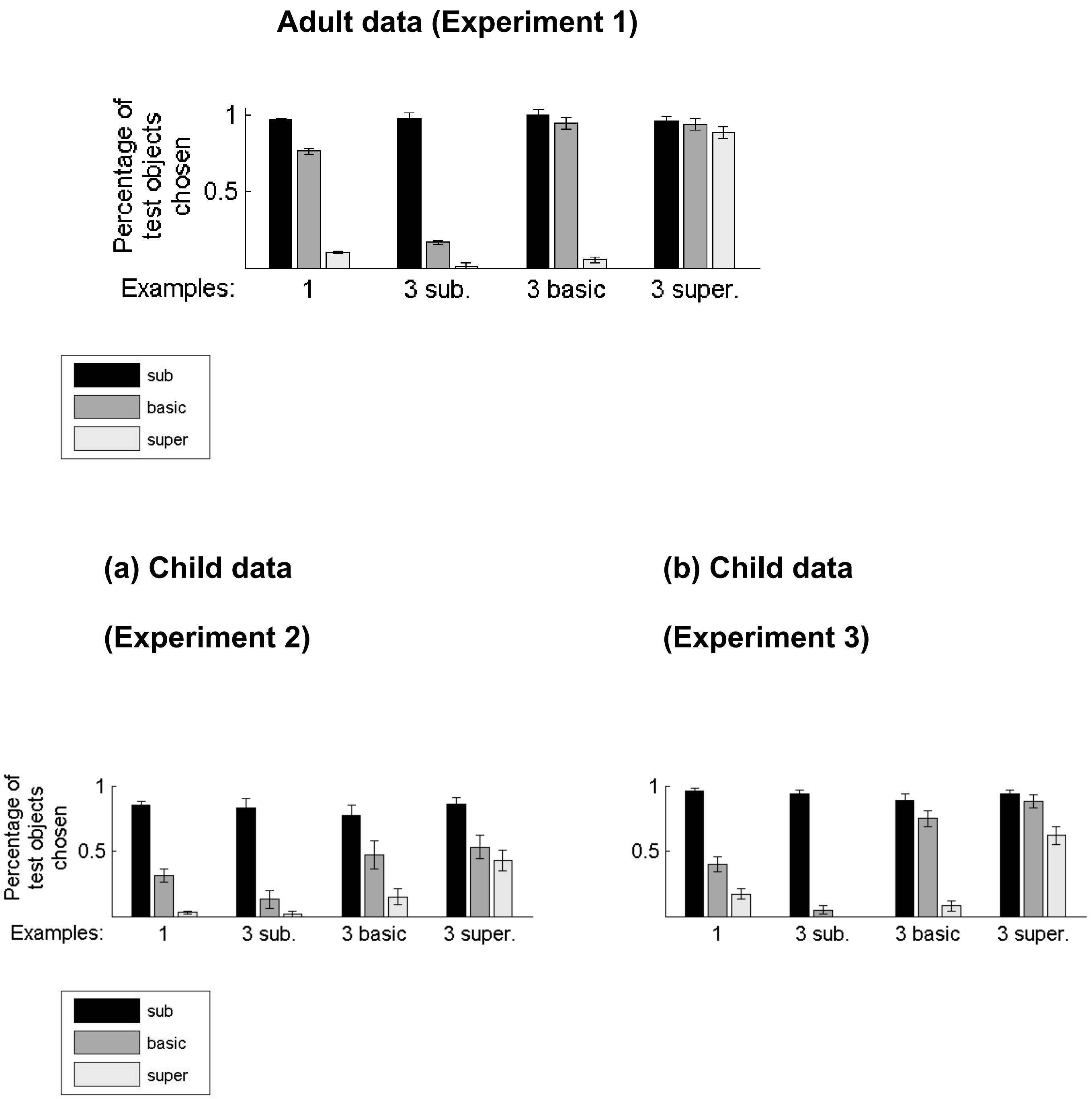
Figure 1. Adults' and children's generalization of word meanings in Experiments 1-3, averaged over domain. Results are shown for each of four types of example set (1 example, 3 subordinate examples, 3 basic-level examples, and 3 superordinate examples). Bar height indicates the frequency with which participants generalized to new objects at various levels. Error bars indicate standard errors.

Figure 2. Predictions of the Bayesian model, both with and without a basic-level bias, compared to the data from adults in Experiment 1 and those from children in Experiment 3.

Figure 3. (a) A schematic illustration of the hypothesis space used to model generalization in the experiment, for the stimuli shown in (b). (b) One set of stimuli used in the experiment, as they were shown to participants.

Figure 4. Percentages of generalization responses at the subordinate and basic levels, for adults and children in both teacher-driven (a) and learner-driven (b) conditions. Corresponding posterior probabilities for subordinate and basic-level hypotheses are shown for the Bayesian model.

Figure 1.



Adult data (Experiment 1)

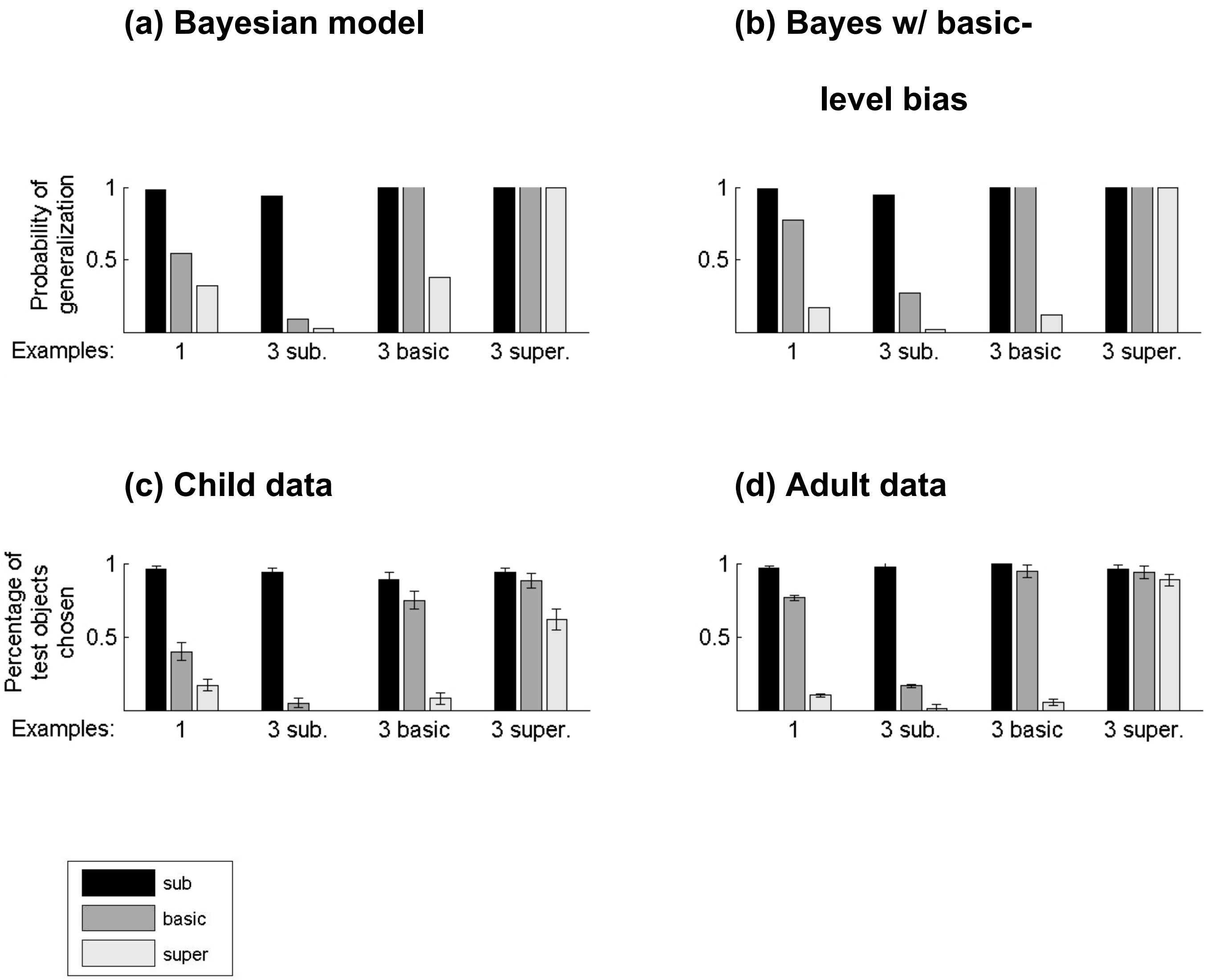
(a) Child data

(Experiment 2)

(b) Child data

(Experiment 3)

Figure 2.



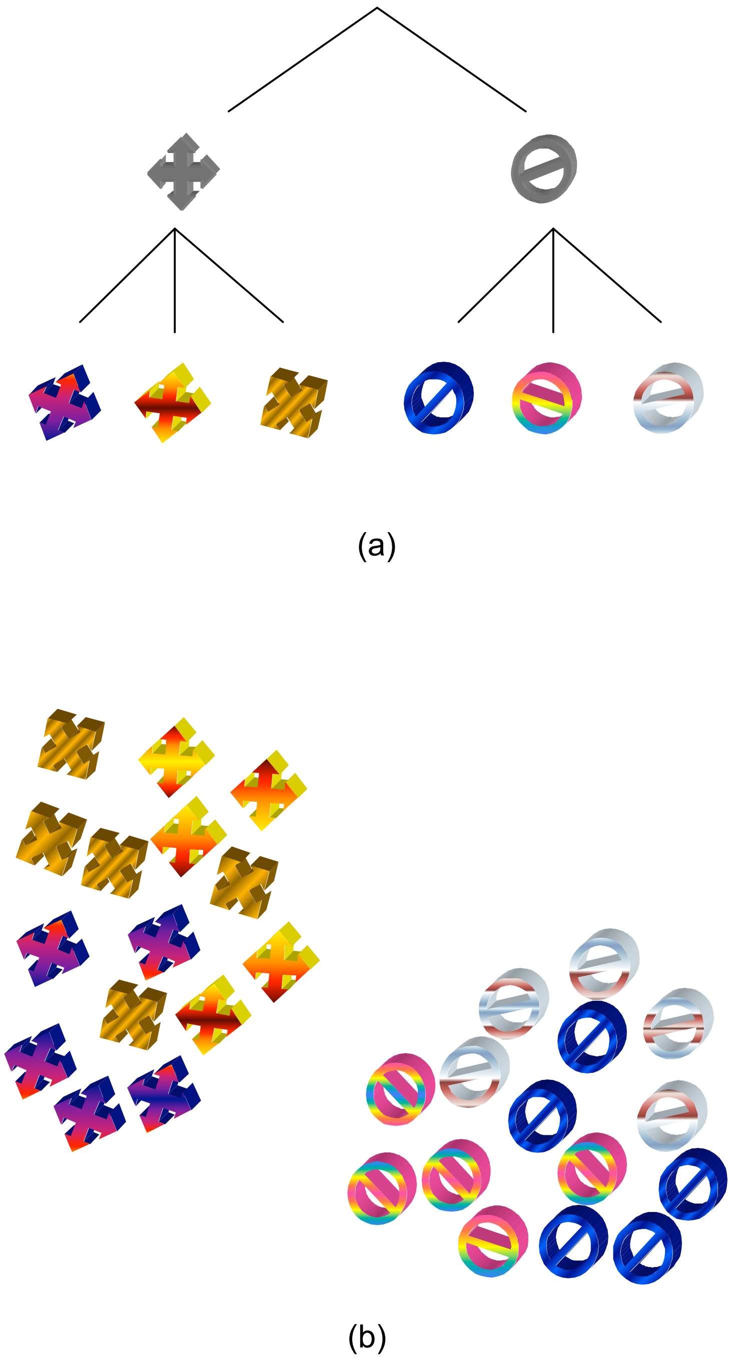
(a) Bayesian model

(b) Bayes w/ basic-level bias

(c) Child data

(d) Adult data

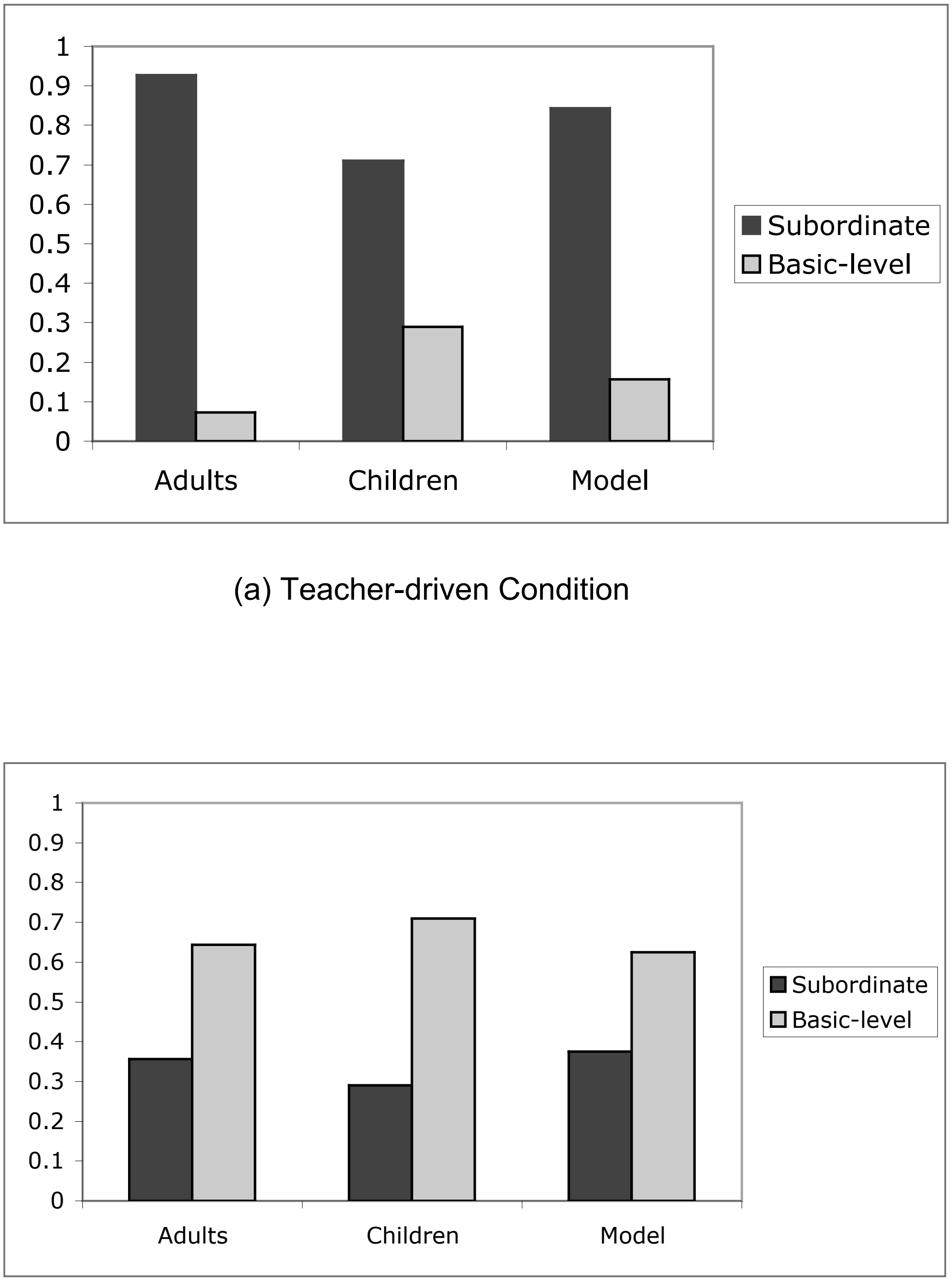
Figure 3.



(a)

(b)

Figure 4.



(a) Teacher-driven Condition

(b) Learner-driven Condition