Infants consider both the sample and the sampling process in inductive generalization

Hyowon Gweon, Joshua B. Tenenbaum, and Laura E. Schulz[1]

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by Susan E. Carey, Harvard University, Cambridge, MA, and approved March 25, 2010 (received for review March 11, 2010)

The ability to make inductive inferences from sparse data is a critical aspect of human learning. However, the properties observed in a sample of evidence depend not only on the true extension of those properties but also on the process by which evidence is sampled. Because neither the property extension nor the sampling process is directly observable, the learner's ability to make accurate generalizations depends on what is known or can be inferred about both variables. In particular, different inferences are licensed if samples are drawn randomly from the whole population (*weak sampling*) than if they are drawn only from the property's extension (*strong sampling*). Given a few positive examples of a concept, only strong sampling supports flexible inferences about how far to generalize as a function of the size and composition of the sample. Here we present a Bayesian model of the joint dependence between observed evidence, the sampling process, and the property extension and test the model behaviorally with human infants (mean age: 15 months). Across five experiments, we show that in the absence of behavioral cues to the sampling process, infants make inferences consistent with the use of strong sampling; given explicit cues to weak or strong sampling, they constrain their inferences accordingly. Finally, consistent with quantitative predictions of the model, we provide suggestive evidence that infants' inferences are graded with respect to the strength of the evidence they observe.

Bayesian model | cognitive development | exploratory play

Human learners can draw rich, abstract inferences from sparse data (1–6). One of the enduring mysteries of cognitive science is why such inferences should be so accurate. The simplest answer is that induction can be accurate as long as the sample is representative of the population. But how do learners know whether a sample is representative? If learners already knew the properties of the population and could see that they were reflected in the sample, they could be confident that the sample was representative. However, it is precisely this information (i.e., the properties of the population) that is in question. Induction is a puzzle because it can hinge on the solution to such chicken-and-egg problems. Inferences about the extension of object properties depend on the relationship between the sample and the population, but knowing that may depend on knowing the extension of the object properties.

The problem of how to infer the extension of object properties from small samples of data bedevils much of scientific inquiry. Rock samples from Mars have a high concentration of silica. Is this true for all Martian rocks or just the (dusty) rocks on the surface? Evergreen needles in a forest lie flat along the branch. Is this true for all needles or only those from low-hanging branches? Scientists could use the appearance of the sample (rocky, needle-like) and/or known category labels ("rocks", "evergreen needles") to generalize properties within but not across kinds (to other rocks and evergreen needles but not from rocks to evergreen needles). Indeed, even young children can use such cues to constrain their inferences (e.g., children infer that entities that share observable properties and/or category labels with a sample are likely to share other properties as well) (7, 8). However, these cues may not suffice. Whether all Martian rocks have silica or all needles lie flat might depend also on the sampling process.

In scientific inquiry, we can usually either control the sampling process or recognize its biases. If, for instance, we *know* that the objects' properties are not independent of the sampling process (because rocks on the surface are more likely to be dusty and to be sampled; because trees low in the canopy have flat needles to maximize sun exposure), we can use this to restrict our generalizations (in both instances, to the population on or near the ground).

However, the problem becomes more complicated when the nature of the sampling process is unknown. This is often the case in social contexts. When a person chooses a sample, she could sample randomly from the whole population or selectively from any subset of the population, for any number of reasons: because of her preferences, because some objects are easier to reach, because she was told what to do, etc. If the person's goals are not made explicit by linguistic or pragmatic cues, the sampling process may not be obvious. Suppose, for instance, a child sees her mother pull a few blue toys from a box of blue and yellow toys. The blue toys squeak. Do all of the toys squeak or just the blue ones? How, short of testing all of the toys, could the child tell?

As in many problems of induction, the problem of generalization from a sample can be solved either by assuming more constraints on the learner, allowing for relatively simple inferences, or by assuming fewer constraints and more sophisticated inferential abilities. Thus one possibility is that there are early constraints on what infants assume about agents' sampling processes. Infants might, for instance, assume *weak sampling* (i.e., agents choose items at random from the population, independent of the properties they have) or *strong sampling* (agents sample items selectively, depending on the properties they have) (9). Alternatively, infants might not have expectations about sampling processes; rather, they might simultaneously infer both the sampling process and the extension of object properties from data. That is, infants might make joint inferences about the subset of the population that was sampled and the subset to which the property extends, given both the possibility that the subset sampled might be independent of the property's extension and the possibility that it might be coextensive with it.

Whether assumed or inferred, the key question is whether infants consider the sampling process and use it to make accurate generalizations. As the names indicate, weak sampling is a less powerful constraint on induction than strong sampling (9). If the learner thinks the evidence was sampled from the population as a whole, then both positive and negative evidence (these toys squeak; those toys do not) is needed to constrain inferences to subpopulations (only this kind of toy squeaks). By contrast, under the strong sampling assumption, even a few samples of positive evidence (these toys squeak) can constrain inductive generalizations to subpopulations or kinds (only this kind of toy squeaks). Here we

Author contributions: H.G. and L.E.S. designed research; H.G. performed research; J.B.T. contributed The Bayesian model; H.G. analyzed data; and H.G., J.B.T., and L.E.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: lschulz@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/1003095107/DCSupplemental.

propose a formal model that captures the relationship between the sampling process, the observed data, and the extension of object properties. We present evidence suggesting that infants can flexibly constrain their predictions about the extension of an object property given the assumed, or inferred, sampling process. In particular, we show that in the absence of behavioral cues to the sampling process, infants make inferences consistent with the use of strong sampling. Critically, this is not because infants cannot consider other alternatives; given explicit behavioral cues to weak or strong sampling, infants constrain their generalizations accordingly.

Our studies build on previous work suggesting that infants may be sensitive to each component of the problem in isolation. That is, infants are capable both of inductive generalization and of sensitivity to sampling processes. Young children project properties across entities that share labels and/or perceptual features (7, 8), and infants as young as 9 months can generalize otherwise hidden properties of objects (e.g., rattling, squeaking) to identical-looking objects after a single exposure (10). Infants in their first year can form expectations about the properties of a sample from a population and about a population from a sample (11), and these expectations are sensitive to how samples are generated: 11-month olds expect randomly generated samples to be representative of the population from which they are drawn, but suspend this inference if the sample is clearly generated selectively (e.g., by an experimenter who expresses a preference for particular objects) (12). Older children can make analogous inferences in reverse. They assume that an agent who pulls a nonrepresentative sample from a population must have a preference for members of that sample but they do not make this inference if the agent pulls a representative sample (13). Finally, the scope of preschoolers' generalizations about word meanings has been shown to depend on both the sample of evidence provided and the nature of the sampling process, in ways predicted by rational Bayesian models of generalization (14, 15). Given three labeled examples of a novel object category, preschoolers restricted their generalizations about the label to the tightest category containing the examples, but only when given explicit cues that the examples were generated by strong sampling rather than weak sampling.

Collectively, these results suggest that infants can project properties from samples to populations, recognize when samples are and are not representative of target populations, and recognize that different sampling processes generate different samples. However, in most previous work the sampling process was specified by explicit social/pragmatic cues (e.g., choosing blindfolded vs. choosing with open eyes and smiling at the chosen items). No previous work has looked at what inferences infants draw when the sampling process is not explicitly cued. Moreover, no previous work has looked at whether infants' generalization of object properties depends on the sampling process. What happens when the probability of drawing a sample and the determination of objects' properties mutually constrain one another? Do infants vary their inferences depending on the relationship between the sample and the population? And can they modulate their generalizations in proportion to how much evidence they have?

Both our model and our experiment follow from the toy box example we outlined above. In the current study, we vary the ratio of blue to yellow balls in a box and the number of blue balls the experimenter pulls from the box. The experimenter squeezes each blue ball in the sample so it squeaks. In all conditions, the question is whether, consistent with different compositions of the sample relative to the population, infants will generalize the squeaking property to the yellow balls. Because the infancy research suggests that babies have abilities presumably prerequisite to such inferences (property projection and sensitivity to sampling processes) by the end of the first year, we look to the beginning of the second year (mean: 15 months) for children's ability to use information about the sample and population to constrain their inferences about the property extension.

Behavioral Study and Comparison with Model Predictions

Our predictions are informed by a Bayesian inference model that formalizes the claim that inductive inferences about object properties depend on both the sampling process (S) and the true extension of the object properties (T). This joint dependence can be described in terms of a simple graphical model (Fig. 1). For simplicity, we consider just three possible property extensions (t_1, the property applies only to blue balls; t_2, it applies only to yellow balls; and t_3, it applies to all balls) and two possible sampling processes (s_1, selectively sampling from just the squeaking set of balls, or strong sampling; s_2, randomly sampling from the whole box, or weak sampling).* The learner observes data $D = n$ examples of blue balls that squeak, drawn from a box that appears to contain a fraction β of blue balls and $1 - \beta$ yellow balls. The learner's goal is to predict Y, the proposition that yellow balls squeak. Note that Y depends directly on T, not S or D; given that we know the set of balls that squeak, the observed data or the process by which the data were sampled are irrelevant to predicting whether the yellow balls squeak. However, inferences about T from D must take into account the different possible values of S. (See *SI Text* for a formal description.) Because the learner's data are inconsistent with the hypothesis that only yellow balls squeak (t_2), only two hypotheses for T are relevant to Y and they make opposite predictions: t_1 predicts that yellow balls do not squeak; t_3 predicts that they do. The output of our model is a likelihood ratio

$$L = \frac{P(D|t_3)}{P(D|t_1)} = \frac{P(n|t_3, \beta)}{P(n|t_1, \beta)}$$

capturing the strength of evidence (16) for t_3 over t_1 and thus for whether yellow balls squeak, while taking S into account. The higher L is, the more likely that yellow balls squeak. In our experiments, we assume that children's exploratory behavior (i.e., how much they squeeze the yellow ball, expecting a squeak) will be monotonically related to L.

As explained in *SI Text*, the likelihood ratio can be expressed as

$$L = \frac{\beta^n}{\alpha + \beta^n(1-\alpha)},$$

where the parameter α describes the learner's prior probability (degree of belief independent of the data D) for selective (or strong) sampling ($S = s_1$). By setting this parameter appropriately, the model can express different possibilities for how infants might take into account sampling in their inductive generalizations. Setting α to either 0 or 1 encodes a definite assumption about the sampling process; setting $\alpha = 0.5$ means that the learner has no initial bias for either sampling process and must make a joint inference about sampling and the property's extension from the observed data.

In our behavioral experiments (see Fig. 2), infants (mean, 15 months, 15 days; range, 13–18 months) saw an experimenter draw blue balls from a box and were then given the inert yellow ball. †In Exps. 1–3, we varied the number of balls drawn from the box (n) and the ratio of blue to yellow balls in the box (β) to provide a sample of balls that was either probable or not probable given the population. Because there is no evidence that infants have initial expectations about the sampling process, we present our data with

*It is possible to generate more complex hypotheses for both the sampling process (*Discussion*) and the property extension (e.g., the three blue balls in the sample plus one other ball might squeak, the three balls in the sample plus two other balls might squeak, etc.). Here we model the simplest set of hypotheses needed to explain the range of evidence presented to infants across all five experiments.

†The model mirrors the task design in distinguishing the sampling phase from the test phase. Because the yellow ball was treated differently from the blue ball(s) (i.e., given directly to the children and not manipulated by the experimenter), we do not treat the yellow ball as part of the sample in the model.

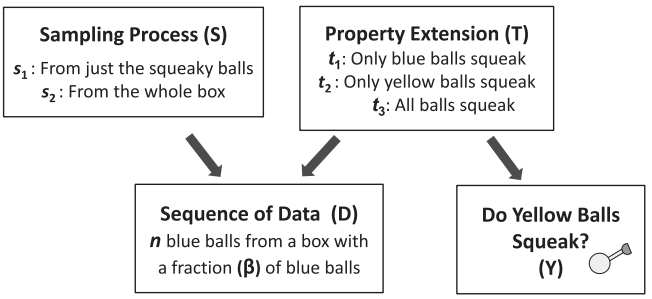

Fig. 1. Graphical model of the relationship between the sample, the sampling process, and the true extension of the object properties. Whether the yellow balls squeak depends only on the extension of the target property. However, the property extension can be inferred only from the observable data (the ratio β of blue/yellow balls in the box and the *n* in the sample), which depend also on the sampling process. Thus to decide whether the yellow ball will squeak, children must either assume a particular sampling process or make a joint inference about both the sampling process and property extension.

respect to the joint inference account (α = 0.5). We then discuss the relationship of the data to the model predictions under definite assumptions of either random (weak) or selective (strong) sampling (α = 0 or 1, respectively). In Exps. 4 and 5, we provide behavioral cues suggesting that the balls are sampled randomly (Exp. 4) or selectively (Exp. 5) to look at how infants' inferences are affected by explicit evidence about the sampling processes. Fig. 3 shows the different strengths of evidence (L) predicted by our Bayesian analysis in these different experimental conditions.

In Exp. 1, 15 children were randomly assigned to a Blue3balls condition and 15 to a Yellow3balls condition. (The color refers to the majority of objects in the box and the number to the number of blue balls drawn.) In both conditions, children saw a box with a transparent front. In the Blue3balls condition, 12 blue balls and 4 yellow balls were visible (β = 0.75); in the Yellow3balls condition, 12 yellow and 4 blue balls were visible (β = 0.25). In both conditions, the experimenter took three blue balls from the box, one at a time. Each time she said "Look!", squeezed the ball so that it squeaked, and then set it on the table. Her actions were identical across conditions, so there were no cues to indicate whether she was sampling from a specific subset of balls or from all of the balls. In both conditions, the experimenter then removed an (inert) yellow ball from the box and gave it to the child. The child was

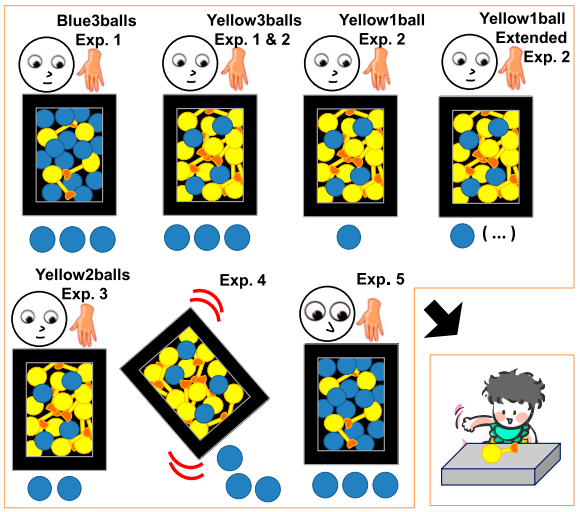

Fig. 2. Schematic of design in Exps. 1–5. See text for details.

allowed to play with the yellow ball for 30 sec. We coded the number of children who squeezed the yellow ball and the number of times each child squeezed.

Under the Bayesian framework, children might consider four joint hypotheses about the sampling process and property extension: H1, sampling = squeaking set (s_1), property = blue (t_1); H2, sampling = whole box (s_2), property = blue (t_1); H3, sampling = squeaking set (s_1), property = all (t_3); and H4, sampling = whole box (s_2), property = all (t_3).

In both conditions, three blue balls are removed from the box (n = 3). In the Blue3balls condition, the data (given that three-quarters of the balls in the Blue box are blue; β = 0.75) fail to distinguish the possibility that the experimenter is sampling from only the squeaky balls (s_1) from the possibility that she is randomly sampling from the whole box (s_2). Because the inference about the sampling process is tightly coupled to the inference about the property extension, the data also fail to distinguish the inference that only blue balls squeak (t_1) from the inference that all balls squeak (t_3). Thus all four hypotheses are consistent with the evidence and the status of the yellow toy is unknown. Because the perceptual similarity between the objects supports the property generalization (10), and the statistical data do not weigh against it, we expected children to squeeze the yellow ball.

By contrast, in the Yellow3balls condition, three blue balls (n = 3) are pulled from a box containing only one-quarter blue balls (β = 0.25). The sample is unlikely if the experimenter were randomly sampling from the whole box; it is more probable as a sample from just the squeaky balls. Again, this inference is coupled to the inference about the property extension. Given that the balls were most likely sampled from the squeaky balls, the evidence that three blue balls squeak is more likely under the hypothesis that only the blue balls squeak than under the hypothesis that all balls squeak. Thus the data support inferences s_1 and t_1. The joint hypothesis H1 makes the observed sequence of data more probable than any of the other alternatives. In this condition, children should assume that the yellow ball does not squeak and thus should be unlikely to squeeze it. Assuming that two sampling hypotheses (s_1 and s_2) are equal a priori (α = 0.5), the likelihood ratio (L; *SI Text*) is 0.59 for the Blue3balls condition and 0.03 for the Yellow3balls condition. (See Fig. 3 for model predictions and results throughout.)

The experimental results confirmed the model predictions. Fewer children squeezed the ball in the Yellow3balls than in the Blue3balls condition [33% vs. 80%; $\chi^2(1, n = 30) = 6.65, P < 0.01$] and children squeezed the yellow ball less often [0.87 vs. 2.53; $t(28) = 2.45$, $P < 0.05$]. These results suggest that infants constrained their generalization of the squeaking property to the blue balls in the Yellow3balls but not the Blue3balls condition.

Although the results of Exp. 1 are consistent with our formal analysis, it is possible that children simply assumed that properties true of a member of the majority kind could be generalized to the minority kind, but not vice versa. That is, children might generalize from the blue balls to the yellow ball when most balls were blue (in the Blue3balls condition) but not when most balls were yellow (in the Yellow3balls condition).

Experiment 2 addressed this alternative explanation. We replicated the Yellow3balls condition of Exp. 1 and compared it with a Yellow1ball condition, in which the experimenter drew just one blue ball out of the mostly yellow box. Randomly drawing a single blue ball from a mostly yellow box is not particularly improbable and does not discriminate between s_1 and s_2 or t_1 and t_3. Although the only difference between the two conditions is the number of balls (n) drawn from the box, we expected that children should restrict their generalization of the squeaking property to the blue ball significantly more often in the Yellow3balls than in the Yellow1ball condition.

Of course, when children are shown three blue balls squeaking rather than one, they also see more actions on the blue ball and are exposed to the blue balls for a longer time. Mere added experience

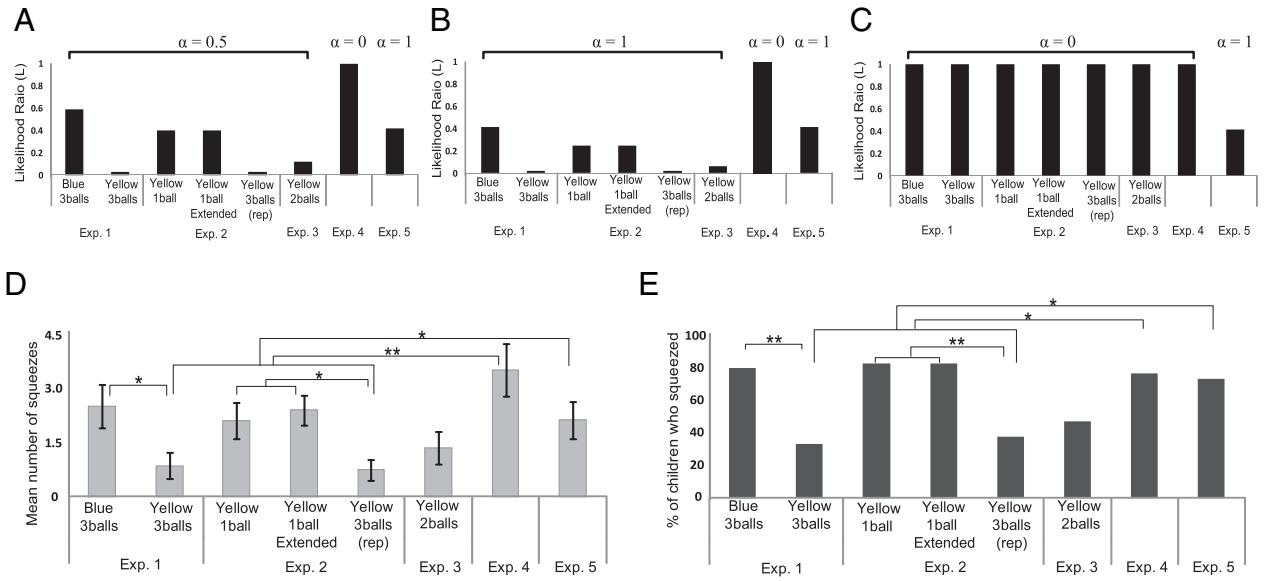

Fig. 3. Model predictions (*A–C*) and results for Exps. 1–5 (*D* and *E*). (*A*) Model predictions with α set to 0.5 (joint inference); (*B*) α set to 1 (assuming strong sampling); (*C*) α set to 0 (assuming weak sampling). In *D* and *E*, asterisks indicate significance in planned comparisons based on model predictions (*, $P < 0.05$; **, $P < 0.01$).

with the blue balls (rather than the number of blue balls in the sample) could make children less likely to generalize the property to the yellow ball. To address this possibility, we ran a Yellow1ball Extended condition in which children again saw a single blue ball drawn from the mostly yellow box. This time, however, the experimenter squeezed the blue ball six times, matching the number of actions and time of exposure to the Yellow3balls condition. If children restrict their generalization to the yellow ball on the basis of the length of exposure and number of actions performed on the blue ball, then children in the Yellow1Ball Extended condition should perform like children in the Yellow3balls condition; if, instead, children are sensitive to the relationship between the sample and the population, children's performance should mirror that of children in the Yellow1ball condition.

With respect to the model, β was held constant (at 0.25) between the conditions whereas n was either 3 or 1. Assuming $\alpha = 0.5$ as in Exp. 1, the likelihood ratio (L) is 0.40 for the Yellow1ball conditions and 0.03 for Yellow3balls replication. Again, the results were consistent with the model predictions. Fewer children squeezed the ball in the Yellow3balls than in the Yellow1ball condition [38% vs. 82%; χ^2 (1, n = 33) = 6.95, $P < 0.01$] and children squeezed less often [0.75 vs. 2.12; $t(31) = 2.35$, $P < 0.05$].

The results of the Yellow3balls condition of Exp. 2 replicated the Yellow3balls condition of Exp. 1 [children squeezing, 38% vs. 33%, P = not significant (NS); mean squeezes, 0.75 vs. 0.87, P = NS] whereas the results of the Yellow1ball condition of Exp. 2 mirrored those of the Blue3balls condition of Exp. 1 (children squeezing, 82% vs. 80%, P = NS; mean squeezes, 2.12 vs. 2.53, P = NS).

These results were not due simply to children's differential exposure to blue balls in the Yellow3balls and Yellow1ball conditions. Children's performance in the Yellow1ball Extended condition was indistinguishable from that of children in the Yellow1ball condition (children squeezing, 82% vs. 82%, P = NS; mean squeezes, 2.41 vs. 2.12, P = NS) and significantly different from children's performance in the Yellow3balls condition. Fewer children squeezed the ball in the Yellow3balls than in the Yellow1ball Extended condition [38% vs. 82%; χ^2 (1, n = 33) = 6.95, $P < 0.01$] and children squeezed less often [0.75 vs. 2.41 ; $t(31) = 2.12$, $P < 0.05$].

These results rule out the alternative explanations discussed above. Although blue balls were the minority objects in both conditions of Exp. 2, children generalized the property in the Yellow1ball condition but not in the Yellow3balls condition. Moreover, although one might assume that the more often infants see an adult squeezing a ball, the more likely they should be to squeeze it themselves, we found the reverse. Infants were more likely to squeeze the yellow ball in the Yellow1ball condition (when the experimenter squeezed only one ball) than in the Yellow3balls condition (when she squeezed three). Whereas this might suggest the other possibility—that the more often infants see an action on a single object kind, the more likely they are to restrict their actions to this kind—this was also ruled out. Infants in the Yellow1ball Extended condition saw the single blue ball squeezed repeatedly but readily generalized the property to the yellow ball. That is, children's tendency to squeeze was unrelated to the number of times they saw the target action but was well predicted by our model in which generalization from the sample depends jointly on the sampling process and the property extension.

Experiment 3 tested the prediction that children's inferences should be graded with respect to the data; that is, children should be progressively less likely to squeeze the yellow ball as the number of balls drawn from the yellow box increases. For instance, setting $\alpha = 0.5$ (s_1 and s_2 are equally likely a priori) and $\beta = 0.25$ (one-quarter of balls in the box are blue), the likelihood ratios (L) are 0.40, 0.12, and 0.03 for $n = 1$, $n = 2$, and $n = 3$, respectively. The significant differences between children's performance in the two Yellow1ball conditions (Exp. 2) and the Yellow3balls conditions (Exps. 1 and 2) provide data for cases in which $n = 1$ and $n = 3$. In Exp. 3 we ran the intermediate case, a Yellow2balls condition, in which the experimenter sampled two blue balls from the box containing one-quarter blue balls. We predicted that children's tendency to squeeze the yellow ball in this condition would be intermediate between the results of two Yellow1ball and Yellow3balls conditions. The prediction of intermediate responding means that although the results of the Yellow2balls condition might not differ significantly from *either* the Yellow1ball or the Yellow3balls conditions, the model estimates for the five conditions should predict the pattern of results. This is what we found. Numerically, more children squeezed in the Yellow2balls condition of Exp. 3 than in either Yellow3balls condition (47% vs. 33%, Exp. 1; 47% vs. 38%, Exp. 2, P = NS) and children squeezed the yellow ball more often (1.35 vs. 0.87, Exp. 1;

1.35 vs. 0.75, Exp. 2; P = NS). Also, fewer children squeezed the yellow ball in either Yellow2balls condition of Exp. 2 [47% vs. 82% (both conditions); χ^2 (1, n = 34) = 4.64, $P < 0.05$], and children squeezed numerically less often [1.35 vs. 2.12 (Yellow1ball), 1.35 vs. 2.41 (Yellow1ball Extended), P = NS]. Critically, this pattern of results was well predicted by the model (Pearson's $r = 0.98$, $P < 0.005$). Given that this correlation considers only five data points, the results should be interpreted with caution. However, they provide suggestive evidence that children's inferences vary in a graded manner with the size of the sample.

In modeling the results so far, we have assumed that the two sampling hypotheses (s_1 and s_2) were assigned equal probability a priori ($\alpha = 0.5$). As noted, infants might instead have initial expectations that agents engage in either weak ($\alpha = 0$) or strong sampling ($\alpha = 1$); we address those possibilities in the discussion to follow. In Exps. 4 and 5, however, we consider the case when children are given overt behavioral cues indicating that the sampling process is either random or selective.

In Exp. 4, the experimenter drew three blue balls from the box with one-quarter blue balls. However, instead of reaching in, she shook the box upside down and let three blue balls fall out. Thus, the evidence [number of balls (n) and proportion of blue balls (β)] was the same as in the Yellow3balls condition of Exps. 1 and 2 but in this case the experimenter's action specified that, despite the improbability of the sample, she was sampling from the whole box. Formally, a direct cue to random sampling sets the parameter α to 0 and raises L to 1.00, much higher than the $L = 0.03$ of the Yellow3balls condition. Thus we predicted that children in Exp. 4 should generalize the squeaking property to the yellow ball more than children in Yellow3balls conditions in Exps. 1 and 2. The results were consistent with this prediction. More children squeezed the ball in Exp. 4 than in the Yellow3balls conditions [76% vs. 33%, Exp. 1; χ^2 (1, n = 32) = 6.03, $P < 0.05$; 76% vs. 38%, Exp. 2; χ^2 (1, n = 33) = 5.13, $P < 0.05$] and children squeezed more often [3.53 vs. 0.87, Exp. 1; $t(30) = 3.24$, $P < 0.005$; 3.53 vs. 0.75, Exp. 2; $t(31) = 3.57$, $P < 0.005$].

What are the predictions in the converse case, when children are given explicit cues to selective sampling but a sample that is also likely under random sampling (three blue balls from the mostly blue box)? We looked at this in Experiment 5. The experimenter reached into the three-quarters blue box but provided cues consistent with selective sampling of a specific set of balls (Fig. 2 and *Methods*). As noted, our model suggests that when β and n are held constant, the likelihood ratio (L) gradually decreases as a function of α. However, the difference in the likelihood between $\alpha = 0.5$ and $\alpha = 1$ is small. With the parameters $\alpha = 1$, $\beta = 0.75$, and $n = 3$, the model predicts only a slightly lower rate of squeezing ($L = 0.42$) in Exp. 5 than in the Blue3balls condition of Exp. 1 ($L = 0.59$) (and thus of course a higher rate of squeezing than in the Yellow3balls conditions of Exps. 1 and 2; $L = 0.03$). Intuitively, this is because explicit cues that the experimenter is selectively sampling from the box (consistent with s_1) do not indicate that the yellow balls are not themselves part of the squeaky set that the experimenter is sampling from; thus the inference that the property extends to the yellow balls continues to depend on β, the ratio of blue and yellow balls in the box. One could, of course, provide social/pragmatic cues that would unambiguously establish that the yellow balls were not being sampled (e.g., by picking the yellow ball, frowning, and replacing it with a blue ball). However, in that context, infants' failure to squeak the yellow ball would be overdetermined (i.e., they could directly infer that the yellow ball should be avoided).

We thus predicted that children in Exp. 5 would generalize the property to the yellow ball, squeezing more than the Yellow3balls condition of Exps. 1 and 2 but no differently from children in the Blue3balls condition of Exp. 1.

The results were consistent with our predictions. There were no differences between the results of Exp. 5 and the Blue3balls condition of Exp. 1 with respect to the number of children squeezing (73% vs. 80%, P = NS) or the mean number of squeezes (2.13 vs. 2.53, P = NS). By contrast, more children squeezed the ball in Exp. 5 than in the Yellow3balls conditions (73% vs. 33%, Exp. 1; χ^2 (1, n = 30) = 4.82, $P < 0.05$; 73% vs. 38%, Exp. 2; χ^2 (1, n = 31) = 4.01, $P < 0.05$), and children squeezed the ball more often [2.13 vs. 0.87, Exp. 1; $t(28) = 2.09$, $P < 0.05$; 2.13 vs. 0.75, Exp. 2; $t(29) = 2.47$, $P < 0.05$].

Thus far we have discussed the joint inference account; we now turn to the possibility that infants might have default assumptions about how agents sample evidence. Our data rule out the possibility that infants assume weak sampling (α fixed to 0). Under the assumption of weak sampling, the model predicts that infants should squeeze the yellow ball persistently in all five experiments (that is, the results of all five conditions should be identical to that of Exp. 4). By contrast, the likelihood ratios under the strong sampling account (α fixed to 1) are quite similar to those under the joint inference account ($\alpha = 0.5$): Exp. 1, Blue3balls condition, $\alpha = 0.5$, $L = 0.59$ vs. $\alpha = 1$, $L = 0.42$; Yellow3balls condition, $\alpha = 0.5$, $L = 0.03$ vs. $\alpha = 1$, $L = 0.02$; Exp. 2, Yellow1ball condition, $\alpha = 0.5$, $L = 0.40$ vs. $\alpha = 1$, $L = 0.25$; Exp. 3, Yellow2balls condition, $\alpha = 0.5$, $L = 0.12$ vs. $\alpha = 1$, $L = 0.06$. Thus our results are consistent with the possibility that infants expect agents to engage in strong sampling. Looking at the overall correlation between the models and our data (mean number of squeezes) across all eight conditions, both the joint inference model and the strong sampling account correlate with the data (joint inference, $r = 0.97$, $P < 0.001$; strong sampling, $r = 0.92$, $P < 0.001$); the weak sampling account does not ($r = -0.07$, P = NS) (Fig. 3).

Given that infants might expect agents to engage in strong sampling, why consider the possibility that they engage in joint inference? As noted, one could make assumptions about infants' prior inductive biases allowing for simpler learning or make no such assumptions and instead credit infants with relatively sophisticated inferential mechanisms. Both the current work (Exp. 4) and previous research (11, 12) establish that infants are sensitive to sampling processes in the presence of explicit behavioral cues. Given that infants recognize that agents can engage in weak sampling and that there is as yet no evidence that infants nonetheless expect agents to engage in strong sampling, joint inference remains a real possibility. That said, considerable work suggests that infants make assumptions about rational agents with respect to intentional goal-directed actions (17–19). It would be very interesting if the assumption that agents were likely to engage in selective sampling were part of this repertoire. Thus distinguishing the strong sampling assumption from the joint inference account remains an important direction for future research.

Discussion

We presented a formal Bayesian account of how inferences about the extension of object properties from a sample of evidence depend on both the true extension of the property and the sampling process. We showed empirically that, given identical samples of evidence, 15-month-old infants make different inferences about the extension of object properties depending on the probability of the sample. In particular, we showed that in the absence of behavioral cues to the sampling process, infants draw inferences consistent with the use of strong sampling; infants were able to draw normative, flexible inferences about the extension of an object property given only a small sample of positive evidence or the property. Additionally, we showed that infants recognize that agents can engage in different sampling processes; given behavioral cues to either weak or strong sampling, infants varied their inferences accordingly. Across the eight conditions, the strength of evidence infants observed for discriminating the two hypotheses about the property extension (all balls squeak vs. only blue balls squeak) predicted their generalizations. Finally, as predicted quantitatively by the Bayesian model, we provided suggestive evidence that infants' inferences are graded with respect to the size of the sample.

We found that both the number of children squeezing and the mean number of squeezes across conditions were consistent with the model predictions. Although the likelihood ratio and these dependent measures were highly correlated, the differences be-

tween the group means in the number of squeezes were mainly driven by the children who did not squeeze at all. Additionally, the all-or-none measure of whether a child squeezed or not showed the same qualitative pattern as the mean number of squeezes. Further computational and empirical research might clarify exactly which aspects of behavior the model predicts.

Throughout, we have looked at the probability that a sample might be randomly generated from the whole population. However, it is possible that children are also sensitive to a different measure of likelihood: the degree to which evidence is *representative* of the population (i.e., the degree to which the evidence in the sample distinguishes the target population from alternative populations). Three blue balls, for instance, may be the most probable draw from a mostly blue box but this sample fails to distinguish a mostly blue box from an entirely blue box. By contrast, a sample consisting of two blue balls and one yellow ball may be a less probable sample but a more representative one (in that it distinguishes the entirely blue from the mostly blue box). The distinction did not arise in the current work because the samples were never distinctively representative (the sample always consisted of only blue balls although the box contained both blue and yellow balls). However, Bayesian inference models can formally capture this distinction (20), and comparing infants' sensitivity to these different measures of likelihood is an intriguing area for future research.

Although we have focused on the distinction between strong and weak sampling assumptions, a variety of more complex models might account for the current data. A child might infer, for instance, that the agent intends to sample squeaky balls and knows which balls squeak, believes that all of the balls squeak, or believes that some balls squeak but does not know which ones. Alternatively, the child might assume that the agent is drawing the sample to teach the child which balls squeak. Recent work in computational modeling has suggested formalizations of both such *intentional* and *pedagogical* sampling assumptions (21, 22). These models make different predictions in a variety of tasks; however, in the current paradigm, the predictions are qualitatively the same. Here we have opted for the simplest model that could explain our data; future research might assess the extent to which infants distinguish more complex sampling assumptions.

Even the current results, however, speak to the sophistication of children's reasoning. These findings suggest that infants make accurate generalizations from sparse data, in part because their inferences are sensitive to how the sample of evidence reflects the population. These results are consistent with the theoretical stance than humans are rational learners from the earliest stages of development. Babies who have just learned to say "mama" and may not yet say "ball" may know something about the goals of the former and infer the properties of the latter simply by attending to the rich statistics of everyday life.

Methods

Behavioral Study. ***Subjects.*** One hundred thirty infants (mean, 15 months, 15 days; range, 13–18 months; 49% girls) were recruited from a local Children's Museum. Fifteen participants were replaced due to (*i*) fussing out, (*ii*) refusal to touch the stimuli, or (*iii*) parental interference. Three additional infants were replaced due to experimental error; one infant in the Yellow3balls condition in Exp. 2 was an outlier, squeezing the ball 3 standard deviations more than the mean, and was excluded from subsequent analysis.

Materials. Two foam-board boxes were used ($30 \times 45 \times 30$ cm). One contained 12 blue and 4 yellow balls; the other contained 4 blue and 12 yellow balls. All balls were visible through a window in the front of the boxes. The top of each box had a hidden compartment from which the target balls could be pulled (Exps. 1–3 and 5) or poured (Exp. 4), without changing the view through the window. The yellow and blue balls were perceptually similar but the yellow balls had wooden handles, providing a "banging" affordance so the child could readily engage in a behavior other than squeezing. The blue balls squeaked; the yellow balls were inert.

Procedure. Children were tested individually. A box sat on a table in front of the child, covered with a cloth. The experimenter revealed the box and drew the child's attention to its contents by pointing to the window. In Exps. 1–3, the experimenter glanced into the box, pulled out a blue ball, squeezed it so that it squeaked, and then set it on the table. She repeated this until she pulled out the number of blue balls for the condition. The experimenter paused, then pulled out a yellow ball and put it in front of the child saying, "Here you go, you can go ahead and play." If the child did not touch the ball, she encouraged the child again. The child was allowed to play for 30 sec. In Exp. 4, rather than pulling the balls out, the experimenter shook the box upside down to let the balls fall out. Then she told the child, "The next one is going to be yours." This comment was added to prevent the infants from anticipating that they would get the box to shake (rather than the balls to squeeze). She shook the box again to let a yellow ball fall out and gave it to the child. In Exp. 5, the experimenter peered into the box and took approximately twice as long as in Exps. 1–3 to pull each blue ball out. As she took the blue ball out, she said, "Aha, here it is, look!" and smiled. After three balls were removed, she said, "The next one is going to be yours" and shook the yellow ball out (matching Exp. 4). In all conditions, children's actions during the 30 sec of play were coded. An additional coder, blind to condition, recoded all data. Intercoder reliability averaged 94.7%. Parents provided informed consent; the MIT Institutional Review Board approved the research.

ACKNOWLEDGMENTS. We thank Phoebe Neel and Camille Doykan for help with data collection, Sydney Katz for help with blind coding, and members of the Early Childhood Cognition Lab for helpful comments and suggestions. This research was supported by a National Science Foundation Faculty Early Career Development Award, by a John Templeton Foundation Award (to L. E.S.), and by a James S. McDonnell Foundation Collaborative Interdisciplinary Grant on Causal Reasoning (to L.E.S. and J.B.T.).

1. Carey S (2009) *The Origin of Concepts* (Oxford Univ Press, New York).
2. Wellman H, Gelman S (1992) Cognitive development: Foundational theories of core domains. *Annu Rev Psychol* 43:337–375.
3. Gopnik A, et al. (2004) A theory of causal learning in children: Causal maps and Bayes nets. *Psychol Rev* 111:3–32.
4. Keil F (1989) *Concepts, Kinds, and Conceptual Development* (MIT Press, Cambridge, MA).
5. Carey S (1985) *Conceptual Change in Childhood* (MIT Press, Cambridge, MA).
6. Schulz LE, Goodman ND, Tenenbaum JB, Jenkins AC (2008) Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition* 109:211–223.
7. Gelman S, Markman EM (1986) Categories and induction in young children. *Cognition* 23:183–209.
8. Gopnik A, Sobel D (2000) Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Dev* 75:1205–1222.
9. Tenenbaum JB (1999) A Bayesian framework for concept learning. PhD thesis (MIT, Cambridge, MA).
10. Baldwin DA, Markman EM, Melartin RL (1993) Infants' ability to draw inferences about nonobvious object properties: Evidence from exploratory play. *Child Dev* 64:711–728.
11. Xu F, Garcia V (2008) Intuitive statistics by 8-month-old infants. *Proc Natl Acad Sci USA* 105:5012–5015.
12. Xu F, Denison S (2009) Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition* 112:97–104.
13. Kushnir T, Xu F, Wellman HM (2008) Preschoolers use statistical sampling information to infer the preferences of others. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, eds Love BC, McRae K, Sloutsky VM (Cognitive Science Society, Austin, TX), pp 1563–1566.
14. Xu F, Tenenbaum JB (2007) Word learning as Bayesian inference. *Psychol Rev* Vol 114, pp 245–272.
15. Xu F, Tenenbaum JB (2007) Sensitivity to sampling in Bayesian word learning. *Dev Sci* 10:288–297.
16. Tenenbaum JB, Griffiths TL (2001) Generalization, similarity, and Bayesian inference. *Behav Brain Sci* 24:629–640.
17. Gergely G, Nádasdy Z, Csibra G, Bíró S (1995) Taking the intentional stance at 12 months of age. *Cognition* 56:165–193.
18. Gergely G, Csibra G (2003) Teleological reasoning in infancy: The naive theory of rational action. *Trends Cogn Sci* 7:287–292.
19. Woodward AL (1998) Infants selectively encode the goal object of an actor's reach. *Cognition* 69:1–34.
20. Tenenbaum JB, Griffiths TL (2001) The rational bases of representativeness. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society,* pp 1036–1041.
21. Goodman ND, Baker CL, Tenenbaum JB (2009) Cause and intent: Social reasoning in causal learning. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, eds Taatgen N, van Rijn H (Cognitive Science Society, Amsterdam), pp 2759–2764.
22. Shafto P, Goodman ND (2008) Teaching games: Statistical sampling assumptions for pedagogical situations. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, eds Love BC, McRae K, Sloutsky VM (Cognitive Science Society, Austin, TX), pp 1632–1637.