Supplementary materials: *A pragmatic theory of generic language*

Michael Henry Tessler (mtessler@stanford.edu)
Noah D. Goodman (ngoodman@stanford.edu)

October 8, 2015

A Experiment 1a: *Measuring the prevalence prior for familiar categories*

The goal of this experiment was to measure participants' beliefs about the distributions of the prevalence of various properties.

A.1 Participants

We recruited 60 participants over Amazon's crowd-sourcing platform Mechanical Turk (MTurk). Participants were restricted to those with US IP addresses and with at least a 95% MTurk work approval rating. 3 participants where unintentionally allowed to do the experiment for a second time; we excluded their second responses (resulting in $n = 57$). 2 participants self-reported a native language other than English; removing their data ($n = 55$) has no effect on the results reported. The experiment took about 10 minutes and participants were compensated \$1.00.

A.2 Procedure and materials

On each trial of the experiment, participants filled out a table where each row was an animal category and each column was a property[1].

Participants were told

> In this study, we are interested in how prevalent certain properties are within different kinds of animals. We will give you examples of the kinds of animals we have in mind and ask you to list a few of your own.
>
> Then, you will estimate the *percentage of the individual members* of the animal species that have certain properties.
>
> On each trial, you will rate 8 properties. The properties will be revealed to you one at a time. Essentially, you will be filling out a big table. You are allowed to go back and revise your answers, if you think there is a more realistic estimate you could give. You will do this 2 times (2 big tables).

The trial proceeded as follows: Participants were first shown a column of 6 pre-specified names of kinds of animals (chosen randomly from a larger set). They were asked to add five of their own animal names to the column. Upon completion, a column appeared with a property as the column header (e.g. "are female"). Participants were asked to fill in each row with the percentage of members of each of the species that had the property (e.g. "50%"). The pre-specified animal categories were those that would be used in the generic truth judgment task (section B). This procedure was repeated for 8 properties in total. The animal kind column was unable to be modified once completed. The prevalence columns, however, could be modified at a later stage (i.e. the participant could go back and change her estimates after seeing more properties). Participants completed this procedure on each of 2 trials for a total of 16 properties per participant.

We used a set of properties associated with generics of theoretical interest (21 properties in total), motivated in part by conceptual distinctions pointed out by Prasada, Khemlani, Leslie, and Glucksberg (2013). The conceptual categories used to generate the generics were: majority characteristic (e.g. *Leopards have spots*), minority characteristic (e.g. *Lions have manes*), striking (e.g. *Sharks attack swimmers.*), and majority false generalizations (e.g. *Robins are female.*) (Prasada et al., 2013). In addition, we included false sentences (e.g. *Lions lay eggs.*), to cover the full range of possible truth values. We aimed to include properties whose generics would be judged false even though the majority of the species has the property. To achieve this, 5 of the properties were negations (e.g. DOESN'T ATTACK SWIMMERS) and were not measured separately from the corresponding positive form. Instead, we took the complement of participants judgments of the positive property (e.g. HAS BEAUTIFUL FEATHERS) as the estimate for the negation (e.g. DOESN'T HAVE BEAUTIFUL FEATHERS).

For a full list of the properties, and generic sentences used in the next experiment, see Table 2.

A.3 Bayesian data analysis

We will describe the analysis model from the perspective of inferring the distribution of a particular property (e.g. LAYS EGGS). The same general structure is implemented for each item independently.

We hypothesized that the prior elicitation task has an important latent structure. Participants are asked to report the prevalence of the property that they believe lies within different kinds of animals. If participants believe that some kinds have the potential to have the property, while others do not, then we would expect $P(x)$ to be structured as a mixture distribution (Griffiths & Tenenbaum, 2005). The mixture is between kinds that can have the property and kinds that cannot, and the weighting assigned to each mixture component is governed by the parameter θ.

[1] The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/experiments/real-kinds/prior-2.html`

We assume that kinds that *cannot* have the property can only have 0% prevalence. Thus, the number of 0% responses in the data ($d_{observed} = 0$) act as an (inverse) indicator of the property's potential to be present in a category θ (for example, the property IS FEMALE has a very high potential to be in a present in a category, whereas the property LAYS EGGS has relatively less potential e.g. LIONS don't have the potential to lay eggs, while they do have the potential to be female). We assume that kinds that can have the property have the property's prevalence distributed according to a Beta distribution with mean γ and concentration ξ, while kinds that cannot have the property can only have 0% prevalence [2].

$$\begin{aligned} d_{observed} &\sim \begin{cases} \text{Beta}(\gamma, \xi) & \text{if Bernoulli}(\theta) = \textsc{t} \\ \delta_{x=0} & \text{if Bernoulli}(\theta) = \textsc{f} \end{cases} \\ &\sim \theta \cdot \beta(\gamma, \xi) + (1 - \theta) \cdot \delta_{x=0} \end{aligned}$$

We put uninformative priors over the mixture parameter θ, and the mean and concentration of the Beta distribution γ and ξ.

$$\begin{aligned} \theta &\sim \text{Uniform}(0, 1) \\ \gamma &\sim \text{Uniform}(0, 1) \\ \xi &\sim \text{Uniform}(0, 50) \end{aligned}$$

We implemented this Bayesian statistical model using the probabilisitic programming language WebPPL (Goodman & Stuhlmüller, 2014). We used the Metropolis-Hastings algorithm to estimate the posterior distribution over the latent parameters θ, γ, and ϕ with 3 MCMC chains of 100,000 iterations removing the first 50,000 iterations.

We also estimated the likely prevalence of each animal category and property pair of interest (e.g. the prevalence of LAYS EGGS among ROBINS). For this, we assume the data was generated from a Beta distribution with unknown mean and concentration.

$$\begin{aligned} \gamma_{kind::property} &\sim \text{Uniform}(0, 1) \\ \xi_{kind::property} &\sim \text{Uniform}(0, 50) \\ d_{kind::property} &\sim \text{Beta}(\gamma_{kind::property}, \xi_{kind::property}) \end{aligned}$$

A.3.1 Prior model validation

To see if our structured, statistical model of the prior elicitation task matched the observed data (and thus, validate our assumptions about the structure of the prevalence prior), we reconstructed likely distributions of prevalence by taking samples from the posterior predictive distribution of d by running the forward model:

$$x \sim \begin{cases} \text{Beta}(\gamma, \xi) & \text{if Bernoulli}(\theta) = \textsc{t} \\ \delta_{x=0} & \text{if Bernoulli}(\theta) = \textsc{f} \end{cases}$$

These distributions reflect both the property's potential to be present—θ—as well as the distribution of prevalence when the property is present—-Beta(γ, ξ).

We compared these posterior predictive distributions to the prior elicitation data by discretizing the posterior predictive distribution and binning (and normalizing) the counts in the prior elicitation data. We used 12 discrete bins: $\{[0-0.01), (0.01-0.05), (0.05-0.15), (0.15-0.25), ..., (0.75-0.85), (0.85-0.95), (0.95-1]\}$. The posterior predictive distribution of the prior model corresponded well with the prior elicitation data ($r^2 = 0.94$), providing good evidence that our assumption of a structured prior is warranted (Figure 1).

The only substantial disagreement between the prior model and the data occur for the properties IS MALE, IS FEMALE in the bin $(0.45 - 0.55)$. The normalized empirical counts put about 75% probability in this bin; the model predicts roughly 50% probability of this bin. This discrepancy can be attributed to a number of 100% or 0% responses present in the data (e.g. some participants self-generated the category COW, and rated that 100% of cows are female). Since the model assumes a smooth Beta distribution, these responses work to flatten out the distribution.

B Experiment 1b: *Truth judgments of common generics*

The goal of this experiment was to measure participants' beliefs about the acceptability of a number of different generic sentences.

B.1 Participants

We recruited 100 participants over Amazon's crowd-sourcing platform Mechanical Turk (MTurk). Participants were restricted to those with US IP addresses and with at least a 95% MTurk work approval rating. 4 participants were excluded for failing to pass a catch trial. 5 participants self-reported a native language other than English; removing their data has no effect on the results reported. The experiment took about 3 minutes and participants were compensated $0.35.

[2] This modeling approach is similar in spirit to Hurdle Models of epidemiological data where the observed counts of zeros is greater than one would expect from standard models of count data such as the Poisson model (e.g. adverse events to vaccines; Rose, Martin, Wannemuehler, & Plikaytis, 2006).

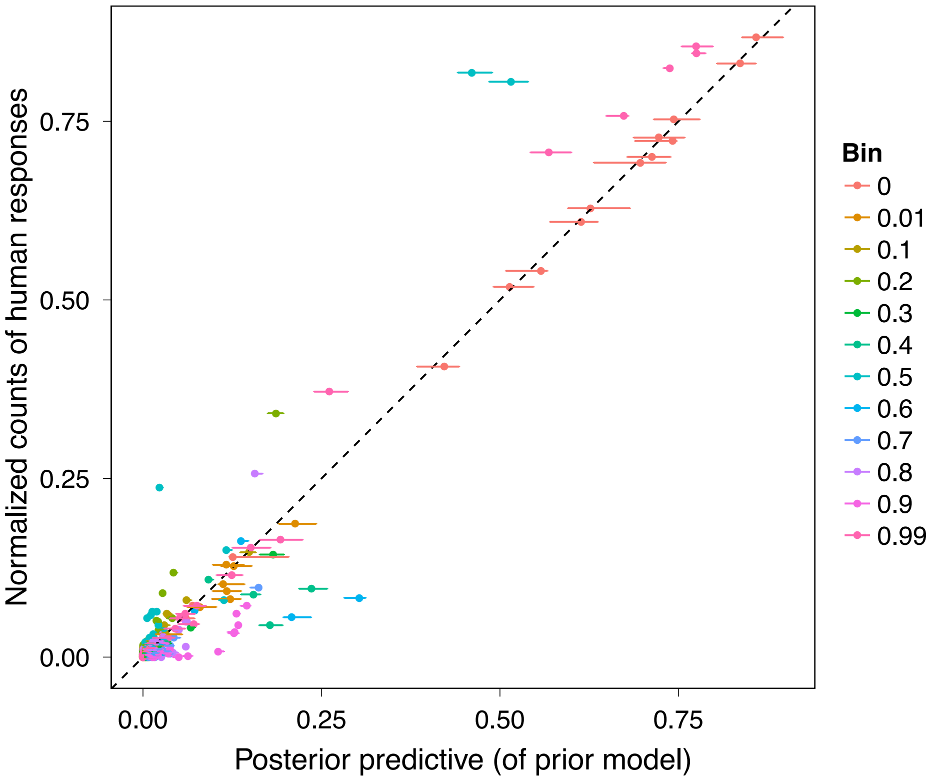

Figure 1: Posterior predictive distribution of the structured, statistical model thought to give rise to the human data in the prior elicitation task. The close alignment between model and data suggests the assumption of a structured prior is warranted.

B.2 Procedure and materials

We used a two-alternative forced-choice (2AFC) paradigm to measure participants' acceptability judgments of 30 generic sentences. Data reported in Prasada et al. (2013) revealed that participants assign midpoint Likert scale ratings to anomalous generics like *Birds are female*. We hypothesized that these "neither true nor false" ratings would be translated into more graded responses across the population using the 2AFC (i.e. subjects would be closer to chance).

Participants were asked to report whether they agreed or disagreed with generic sentences presented in bare plural form (e.g. *Birds lay eggs.*). Items were presented sequentially, and participants reported whether or not they agreed with the sentence by pressing either P or Q (randomized between-subjects).

Generic sentences were selected to correspond with the properties used in Expt. 1a. They covered a range of conceptual categories: Characteristic (e.g. *Ducks have wings.*), Minority (e.g. *Robins lay eggs.*), Striking (e.g. *Mosquitos carry malaria.*), False generalization (e.g. *Robins are female.*), and False (e.g. *Lions lay eggs.*). We also aimed to include generics that were both acceptable and unacceptable with low, medium, and high prevalence. Approximately 10 true, 10 false, and 10 uncertain truth-value generics were selected (see Table 2 for full list of items).

As a manipulation check, participants were asked at the end of the trials which button corresponded to "Agree". 4 participants were excluded for failing this trial.

B.2.1 Results

As a manipulation check, the first author assigned an *a priori* truth-judgment (true/false/indeterminate) to each stimulus item. This was a significant predictor of the empirical truth judgments: true generics were significantly more likely to be agreed with than the indeterminate generics ($\beta = 3.14; SE = 0.15; z = 21.5$), as revealed by a mixed-effect logistic regression with random by-participant effects of intercept. Indeterminate generics were agreed with *less* likely than chance ($\beta = -0.49; SE = 0.09; z = -5.3$) but significantly more than false generics ($\beta = 2.09; SE = 0.14; z = 14.5$).

All results of interest are reported in the main text. Figure 2 shows the human truth judgments as compared to the target-category prevalence of the property ($r^2 = 0.599; MSE = 0.0655$); compare to Figure 2 of the main text. Large deviations from a purely within-kind prevalence account remain: Generics in which the target category had intermediate prevalence (prevalence quartiles 2 and 3: $20\% < prevalence < 64\%$), were not explained at all by prevalence within those categories ($r^2_{Q2,3} = 0.029; MSE = 0.11$).

B.3 Full model predictions

The speaker model S_2 (Eq. 4, main text) assigns a probability to the generic utterance by reasoning about the likely prevalence that a listener (Eq. 1, main text) will infer given each utterance and common sense knowledge of properties (the prior). The model, thus, takes into account the prior knowledge of prevalence distributions; we use the priors inferred from Expt. 1a, maintaining the uncertainty that remains after conditioning on the prior elicitation data.

Having inferred likely priors empirically, the model has 1 parameter: the speaker optimality parameter λ (in Eq. 2, main text). For data analysis, we put an uninformative prior over this parameter, with a range consistent with previous literature using the same model class.

$$\lambda \sim \text{Uniform}(0, 20)$$

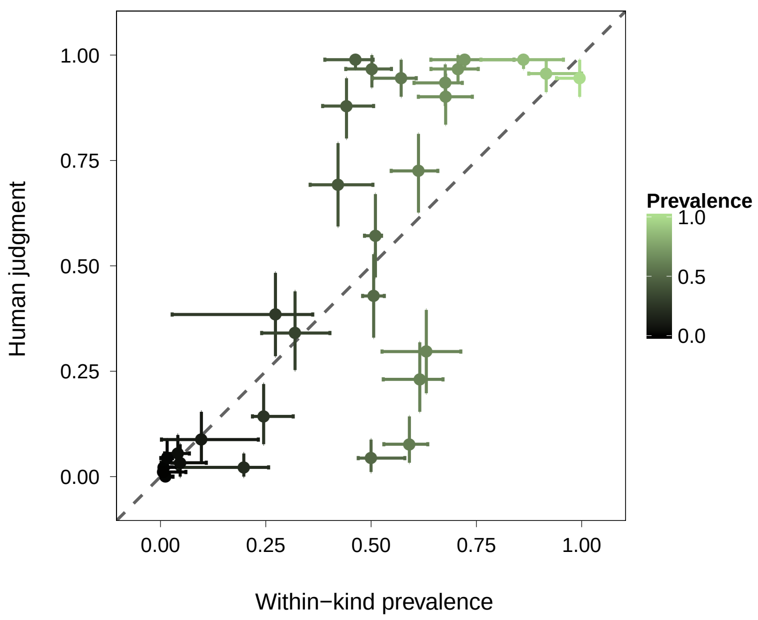

Figure 2: Truth judgments from Expt. 1b for each item vs. the prevalence of the property for the target item as measured in Expt. 1a. For example, "Leopards have spots" has a very high truth judgment (Expt. 1b; Y-axis), and "has spots" is a highly prevalent property for leopards (Expt. 1a; X-axis). Error bars denote 95% confidence intervals for the human judgments and 95% Bayesian credible intervals for the prevalence.

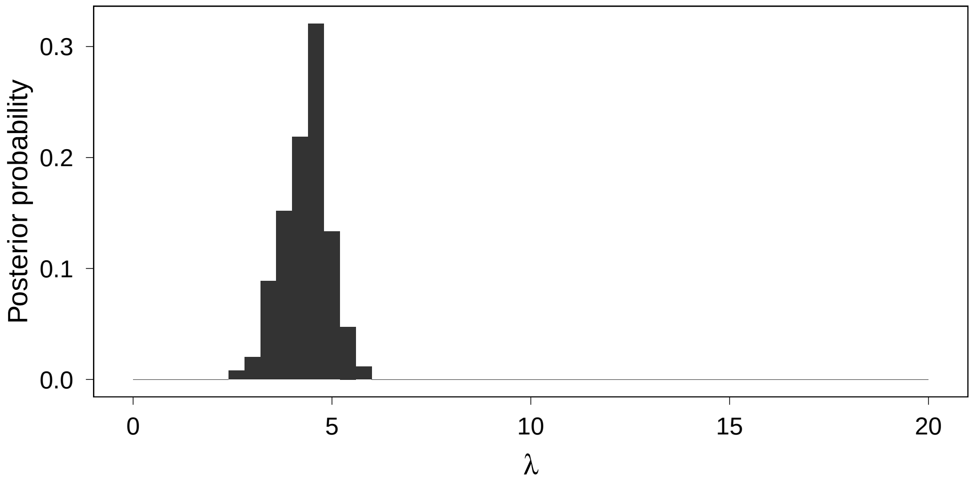

Figure 3: Posterior distribution of the speaker optimality parameter λ in Eq. 2 of the pragmatics model conditioning on the truth judgments of natural cases experimental data. The 95% credible interval is [3.36, 4.98].

In addition, we include a data-analytic "contamination" parameter to account for data that can be reliably be attributed to noise, or random guessing behavior. Modeling random guessing explicitly is important for recovering reliable estimates of the parameters of the model, which would otherwise be contaminated by this data (Lee & Wagenmakers, 2014). The data is thus modeled as a mixture of behavior derived from the S_2 generics model and random guessing behavior. We put an uninformative prior over the mixture parameter ϕ.

$$\phi \sim \text{Uniform}(0,1)$$

B.3.1 Posterior over model parameters

The S_2 generics model has one free parameter, the speaker rationality λ in Equation 2, and one data analytic parameter, a contamination parameter ϕ. To learn about the *a posteriori* credible values of our model parameters, we used the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014) to collect 3 MCMC chains of 100,000 iterations removing the first 50,000 iterations using the Metropolis-Hastings algorithm. The speaker rationality parameter represents the speaker's belief in how rational the hypothetical listener believes he is when choosing to say the generic (over saying nothing). The 95% Highest Probability Density (HPD) Interval is [3.36, 4.98].

In the data analysis, we also include a contamination parameter ϕ to account for noise in the data. This represents the proportion of the data that can be better explained by random guessing than by our model of generic language. In this way, ϕ provides a crude measure of goodness of fit. The 95% HPD Interval is [0.026,0.049]. This suggests that there is not a substantial amount of unexplainable noise in the data.

B.3.2 Posterior predictive

We evaluate the model by examining the posterior predictive distribution of responses. The posterior predictive distribution marginalizes over the inferred parameter values to produce predictions about what the data should look like given the pragmatics model and the observed data. This is akin to fitting the parameters and is an important step in model validation: It shows what data is actually predicted by the model. Figure 5 shows the Maximum A-Posteriori

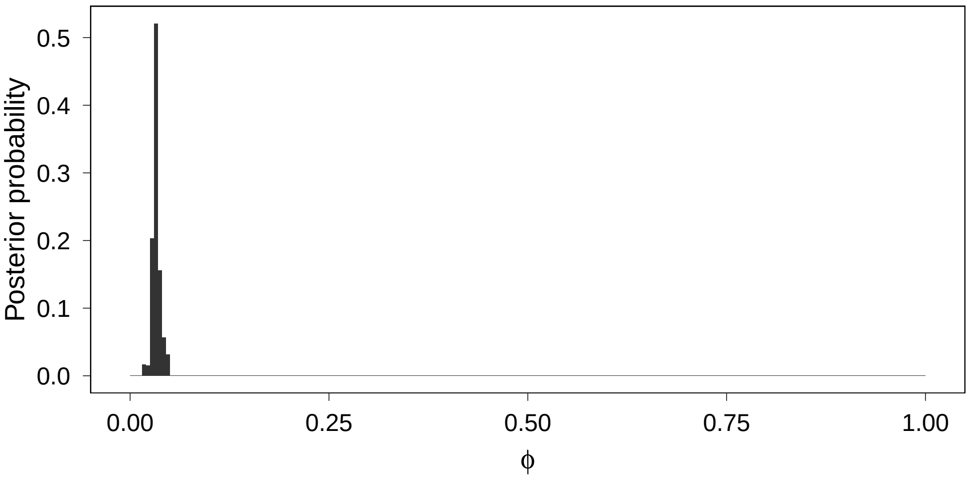

Figure 4: Posterior distribution of the contamination ("guessing") parameter. The 95% credible interval is [0.026,0.049].

(MAP) values of the model predictions compared against the observed data. The model predicts graded endorsements for the generic statements used in Expt. 1b, accounting for almost all of the variance ($r^2 = 0.982; MSE = 0.0035$).

C Experiment 2a: *Measuring prevalence prior for unfamiliar categories*

In this experiment, we measured participants' beliefs about the prevalence distribution of novel properties. We built on the stimulus set from Cimpian, Brandone, and Gelman (2010) which consisted of novel animal categories (e.g. GLIPPETS) and various properties (e.g. HAVE ORANGE LEGS; HAVE BROKEN LEGS).

C.1 Participants

We recruited 40 participants over Amazon's crowd-sourcing platform Mechanical Turk (MTurk). Participants were restricted to those with US IP addresses and with at least a 95% MTurk work approval rating. All participants were native English speakers. The experiment took about 5-7 minutes and participants were compensated $0.75.

C.2 Procedure and materials

To measure the prevalence prior of properties corresponding to *familiar* generics (Expt. 1a), participants filled out a table with rows corresponding to different animal kinds and columns corresponding to different properties. Pilot testing suggested this was a pragmatically strange setup when using novel kinds: answering "What percentage of lorches have green feathers?", when participants knew nothing about lorches, was difficult. Instead, we harnessed the latent structure revealed in prevalence priors by Expt. 1a, described in Section A.3, structuring our task into questions about the property's potential to be present in a kind and the expected prevalence when present; we then used a Bayesian statistical model to reconstruct the underlying prior distribution. We used these distributions in our language model to make predictions about the implications and truth conditions of novel generic sentences.

Participants were told they were on a newly discovered island with lots of new animals on it. They were then given the following instructions:

> One day, you are roaming through the library when you encounter a data-collection robot. The robot doesn't know very much about the world and is asking you questions to learn more. Today, it wants to learn about properties of animals. It is randomly selecting an animal from its memory and a property from its memory, and asking you if the animal is likely to have the property.
>
> Of course, you're new to this island so you don't really know anything about these animals. The properties, however, will be familiar. Try to provide your best guess given your own experience.

Participants were then run through a practice trial where they were familiarized with the questions that would be asked of them. On each trial, the data-collection robot introduced a new animal (e.g. "We recently discovered animals called glippets."). The robot then asked how likely it was that "there was *a* glippet with PROPERTY". Participants responded using slider bars that ranged from "unlikely" to "likely". This question aimed to get at the property's potential to be present (e.g. it's very likely that there is a glippet that is female, less likely that there is a glippet that has wings, and even less likely that there is a glippet that has purple wings). The second question was about the expected prevalence when present. The robot asked, "Suppose there is a glippet that has wings. What percentage of glippets do you think have wings?" Participants responded using slider bars that ranged from "0%" to "100%".

Materials—novel animal names and familiar properties—built upon those from Cimpian et al. (2010). Classic work in generalization suggested to us that there may be differences in the implications of generic statements of different types of biological properties (Nisbett, Krantz, Jepson, & Kunda, 1983). We expanded the stimulus set to include four different types of properties: biological parts (e.g. FEATHERS), colored parts (e.g. GREEN FEATHERS), vague parts (e.g. SMOOTH FEATHERS), and accidental parts (e.g. BROKEN FEATHERS). Pilot testing revealed a lot of variability for items in the accidental properties relative to the other types of properties. To test the quantitative predictive power of the generic interpretation model, we used twice as many exemplars of accidental properties, with the aim to make a "common accidental" and a "rare accidental" class of properties. We used 8 exemplars of each of

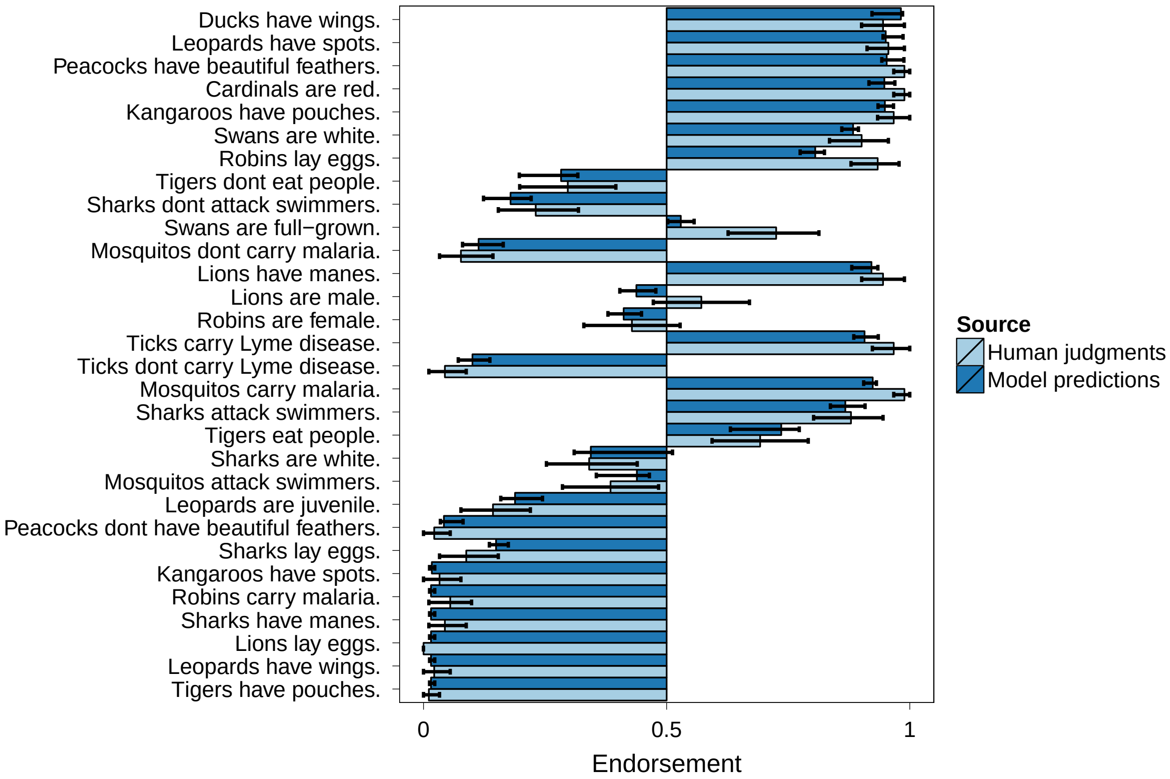

Figure 5: Truth judgments of common generics and model predictions. Items are ordered by within-kind prevalence, and large deviations from a purely within-kind prevalence account are observed for items with intermediate prevalence. The pragmatics model, however, is guided both by the distinctiveness of the property and the *a priori* category prevalence and the communicative pressures to be truthful and informative. The pragmatics model produces a gradient of responses, matching human endorsements well ($r^2 = 0.98$). Error bars denote 95% confidence intervals for the human judgments and 95% Bayesian credible intervals for the model predictions.

the three non-accidental properties (“parts”, “colored parts”, “vague parts”) and 16 exemplars of accidental properties. Materials are shown in Table 3.

C.3 Bayesian data analysis

In order to recover single belief distributions representing both the property’s potential to present in a kind and its expected prevalence when present (analogous to those elicited in Expt. 1a and shown in main text Figure 1), we built a simple Bayesian statistical model of the task questions and their relation to the prevalence distribution of interest.

This analysis was first completed by item to use for predicting (in conjunction with the L_1 model, Eq. 1, main text) interpretation of novel generics. We then performed a median split on the accidental properties type based on the “mean prevalence when present” to construct “rare accidental” and “common accidental” properties types. The same analysis was then completed by property type (e.g. “part”, “color part”, etc...) to test the relationship between interpretation and truth conditions of novel generics (Cimpian et al., 2010).

Participants’ responses to each question (slider bar values) were assumed to be samples from Beta distributions with unknown means and concentrations. The responses to the questions ($d_{potential}, d_{expected}$) were used to estimate the parameters of the distributions of property’s potential to be present in a kind and expected prevalence when present, separately.

$$d_{potential} \sim \text{Beta}(\gamma_{potential}, \xi_{potential})$$
$$d_{expected} \sim \text{Beta}(\gamma_{expected}, \xi_{expected})$$

The parameters γ and ξ correspond to the mean and concentration, respectively, of the distributions over each of these parameters of interest. We put identical, uninformative priors on these parameters:

$$\gamma \sim \text{Uniform}(0,1)$$
$$\xi \sim \text{Uniform}(0,50)$$

To construct single prevalence distributions reflecting both the property’s potential to be present in a kind and the expected prevalence when present (as we have for Expt. 1a), we assume that the distribution is a mixture of categories for which the property is present and categories for which the property is absent. Whether or not a category has a property is driven by the property’s potential to be present (θ).

$$\begin{aligned}\theta &\sim \text{Beta}(\gamma_{potential}, \xi_{potential})\\ x &\sim \begin{cases}\text{Beta}(\gamma_{expected}, \xi_{expected}) & \text{if Bernoulli}(\theta) = \text{T}\\ \delta_{x=0} & \text{if Bernoulli}(\theta) = \text{F}\end{cases}\\ &\sim \theta \cdot \beta(\gamma_{expected}, \xi_{expected}) + (1-\theta)\cdot \delta_{x=0}\end{aligned}$$

Marginal posterior distributions for x were estimated using the Metropolis-Hastings algorithm in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014). Inference was completed by taking 3 MCMC chains of 100,000 samples (removing the first 50,000 samples for burn-in).

D Experiment 2b: *Interpretations of novel generics*

Generics are important for providing information about new or poorly understood categories. As such, it is crucial to look beyond what makes them true and to see how generics are interpreted. Gelman, Star, and Flukes (2002) found that adult and children participants interpreted generics about familiar kinds (e.g. *Bears like to eat ants.*) as implying that all or almost-all bears like to. Cimpian et al. (2010) replicated this conceptually with novel categories (e.g. LORCHES) and found that the strong interpretation of generics could be weakened by predicating the kind with accidental properties (e.g. *Lorches have muddy feathers.*).

The pragmatic listener model L_1 (Eq. 1, main text) predicts that the interpretations of generics should vary as a function of the prevalence prior. Here, we explore whether the predictions based on the empirically elicited prevalence priors for 40 items match human judgments of how the widespread the property is upon hearing a generic. We show how the prevalence-based model predicts the variability in the interpretations of 40 generic utterances with strong quantitative accuracy.

D.1 Participants

We recruited 40 participants over Amazon's crowd-sourcing platform Mechanical Turk (MTurk). Participants were restricted to those with US IP addresses and with at least a 95% MTurk work approval rating. All participants were native English speakers. The experiment took about 5 minutes and participants were compensated $0.60.

D.2 Procedure and materials

In order to get participants motivated to reason about novel animals, they were told they were the resident zoologist of a team of scientists on a recently discovered island with many unknown animals; their task was to provide their expert opinion on questions about these animals[3]. We recruited 40 participants for this *implied prevalence task.*

Participants were supplied with the generic (e.g. *Glippets have yellow fur.*) and asked to judge prevalence: "What percentage of glippets do you think have yellow fur?". Participants saw 25 trials: 5 for each of 5 property types (see Section B.2). The original study by Cimpian et al. found a difference in the implied prevalence between "color parts" (e.g. YELLOW FUR) and accidental properties (e.g. WET FUR). The prevalence priors inferred from Expt. 2a suggest that generic interpretation could be even more variable. For this reason, we included three types of biological properties: parts (e.g. FUR), color–part pairs (e.g. YELLOW FUR) and gradable adjective–part pairs (e.g. CURLY FUR). We also coded the accidental properties from Expt. 2a as either "common" or "rare" using a by-item median split based on *a priori* expected prevalence when present.

Table 1 shows an example trial for two of the five types of property (shown also are the materials for Expt. 2c, truth conditions of novel generics). For a full list of the stimuli used (including examples of the other three types of properties), see Table 3.

		Implied prevalence (Expt. 2b)	Truth conditions (Expt. 2c)
Biological Parts			
	Information	Lorches have green feathers.	xx% of lorches have green feathers.
	Question	What percentage of lorches do you think have green feathers?	Is the following sentence true or false? Lorches have green feathers.
Common Accidental			
	Information	Lorches have muddy feathers.	xx% of lorches have muddy feathers.
	Question	What percentage of lorches do you think have muddy feathers?	Is the following sentence true or false? Lorches have muddy feathers.

Table 1: Sample item from Experiments 2b & 2c

D.3 Full model predictions

The listener model L_1 (Eq. 1, main text) assigns a probability to different levels of prevalence by reasoning about the likely meaning (threshold) of the generic given that the speaker S_1 (Eq. 2, main text) was trying to be informative about the prevalence. The concept of informativity is always with respect to some prior beliefs about the world. As such, the model takes into account the knowledge of the prevalence distribution when interpreting the generic utterance.

D.3.1 Data analysis model priors

As for the task with common generics (Section B.3), we maintain uncertainty about the parameters of the prior in our analysis of the interpretation data. The model, then, has 1 parameter governing the optimality of the hypothetical speaker in Eq. 2. We put an uninformative prior distribution, with a range consistent with previous literature using the same model class: $\lambda_{\text{implied prevalence}} \sim \text{Uniform}(0,20)$.

[3] The experiment in full can be viewed at `http://stanford.edu/~mtessler/generics/experiments/asymmetry/asymmetry-2.html`

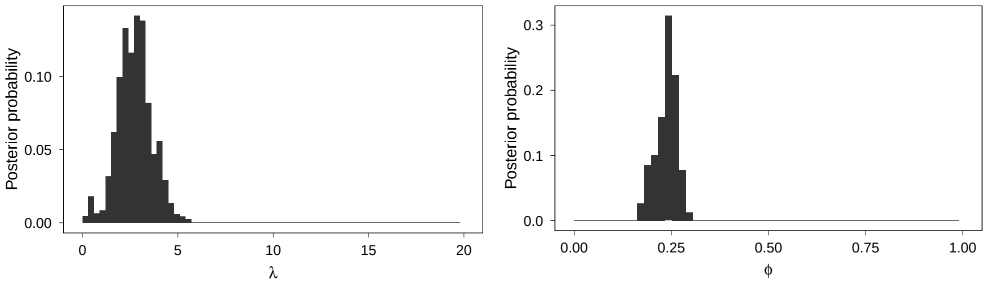

Figure 6: Posterior distribution of the speaker rationality parameter and the contamination parameter for the implied prevalence task.

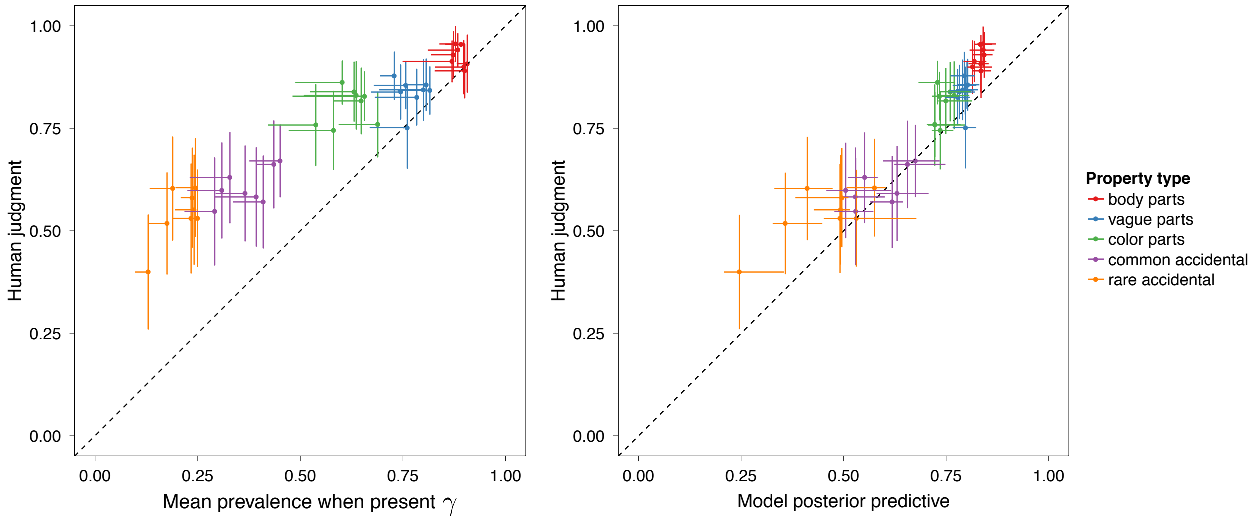

Figure 7: Implied prevalence by human participants vs. *a priori* Mean prevalence when present and the generics model posterior predictive.

As before, we model the observed data as being generated by a mixture of our language model and a model of random guessing behavior. We put an uninformative prior over this mixture parameter $\phi_{\text{implied prevalence}} \sim \text{Uniform}(0,1)$, and infer its credible values from the data.

D.3.2 Posteriors over model parameters

To learn about the *a posteriori* credible values of our model parameters, we used the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014) to collect 2 MCMC chains of 100,000 samples (removing 50,000 for burn-in) using the Metropolis-Hastings algorithm. The estimated posterior distributions of the speaker rationality parameter $\lambda_{\text{implied prevalence}}$ and the contamination parameter $\phi_{\text{implied prevalence}}$ are shown in Figure 6. The contamination parameter represents the proportion of the data that is better explained by a model of random guessing behavior than by the prevalence-based generics model.

The estimated posterior distributions for the two parameters: λ and ϕ are shown in Figure 6. The MAP and 95% credible interval for λ is 3.3 [1.2, 4.7] and ϕ is 0.25 [0.18, 0.28].

D.3.3 Posterior predictive

The posterior predictive distribution compared to human judgments is shown in Figure 7 (Right). As shown in Figure 4 (main text), interpretations of generic utterances are heavily guided by the *a priori* mean prevalence when present. We additionally plot the human judgments (after hearing the novel generic) with the mean prevalence when present (before hearing the generic) in Figure 7 (Left). Though both the posterior predictive distribution of the listener model (Eq. 1) given the generic utterance (7 Right) and the *a priori* mean prevalence when present (7 Left) explain roughly the same amount of variance in generic interpretation ($r^2 = 0.89, 0.92$; MSE = 0.006, 0.042; respectively), the *a priori* mean prevalence when present consistently *underestimates* the interpreted prevalence. This is because the generic implies more than merely that the property is present within the kind; the generic implies the prevalence is substantially more than a listener would expect *a priori*. This is manifested in the pragmatics model by the pressure for the speaker to be informative.

E Experiment 2c: The asymmetry between the implications of novel generics and their truth conditions

Cimpian et al. (2010) observed that generic statements of novel kinds with biological properties (e.g. *Glippets have yellow fur.*) show an asymmetry between the conditions by which the generic is true ("*truth conditions*") and the prevalence implied by the generic ("*implied prevalence*"). Generic sentences were endorsed for a wide-range of prevalence levels (e.g. when "30% of glippets have yellow fur."), resulting in intermediate average truth conditions. As noted above, upon reading a generic, participants inferred that the property was widespread (e.g. almost all glippets have yellow fur). This mismatch between *truth conditions* and *implied prevalence* was significantly reduced for generics of properties plausibly construed as accidental (e.g. *Glippets have wet fur.*). It was also found to not be present for the quantifier "most", arguably the quantifier most similar to a generic. This effect was driven by the sensitivity of the implied prevalence metric; the average truth conditions were modulated by the type of property to a much smaller extent.

Below we replicate the asymmetry findings of Cimpian et al. (2010) and reveal even more variability in the mismatch between *truth conditions* and *implied prevalence* using the types of properties from Expt. 2a. Because of the nature of the truth conditions task, our analysis is by property type (as opposed to individual properties as in Expt. 2b).

E.1 Participants

We recruited 40 participants over Amazon's crowd-sourcing platform Mechanical Turk (MTurk). Participants were restricted to those with US IP addresses and with at least a 95% MTurk work approval rating. All participants were native English speakers. None of the participants completed Expt. 2b (interpretations of novel generics). The experiment took about 5 minutes and participants were compensated $0.60.

E.2 Procedure and materials

The cover story was the same as in Expt. 2b (see Section D.2).

Following Cimpian et al.'s paradigm, in the *truth conditions* task, participants were given a prevalence statement consisting of the percentage of a novel animal category that had a property (e.g. "30% of glippets have yellow fur"). Participants were asked if they agreed or disagreed with the associated generic statement (i.e. *Glippets have yellow fur.*). Prevalence varied between 10, 30, 50, 70, and 90%. The experiment consisted of 25 trials: 5 trials for each of 5 types of properties measured in Expt. 2a (part, color part, vague part, common accidental, rare accidental). Each prevalence level appeared once for each property type (5 prevalence levels x 5 property types). See Table 1 for an example trial.

E.3 Data analysis

To compare the truth judgments data to the implied prevalence data, we computed, for each subject, an *average prevalence level* that led to "Agree" responses, following Cimpian et al. (2010). For example, if a participant agreed with the generic whenever the prevalence was 70% or 90% and disagreed at the other prevalence levels, that participant received an *average prevalence score* of 80%; if a participant disagreed to everything, their *average prevalence score* was 100%, since they presumably would only agree with the generic if the prevalence was 100%.

We subjected our model to the same procedure. Just as with the experiment with common generics, we use the model of the speaker S_2 in Eq. 4 as a model of the *truth conditions* task for novel generics. At each prevalence level, the model returns a posterior probability of saying the generic. We take this probability as the parameter of a Bernoulli trial (i.e. as the weight of a coin), and sample a Bernoulli trial (i.e. flip that coin) to predict whether or not the model would agree to the generic for that prevalence level. We then get a sequence of responses corresponding to the responses (*Agree/Disagree*) a participant would give for the 5 different prevalence levels. Just as with the human data, we took the *Agree* trials, and took the mean of the prevalence levels corresponding to those *Agree* trials. Thus, we computed the average prevalence at which the model assented to the generic for each property type. We repeated this procedure 40 times to simulate a sample of 40 participants. We repeated this procedure 1000 times to bootstrap 95% confidence intervals.

E.4 Full model predictions

We use the implied prevalence data from Expt. 2b, collapsed across property type using the L_1 model to model the *implied prevalence* task. The prior elicitation data (Section C) was analyzed again by property type.

E.4.1 Data analysis model priors

The parameters for the models are the same as for the other tasks: 1 speaker rationality parameter λ and 1 contamination parameter ϕ. Since the nature of the two tasks is quite different (2AFC vs. give a number), we use different parameters for the models (S_2 and L_1). We put uninformative priors over these parameters.

$$\begin{aligned}\lambda_{\text{truth conditions}} &\sim \text{Uniform}(0,20)\\ \lambda_{\text{implied prevalence (by type)}} &\sim \text{Uniform}(0,20)\\ \phi_{\text{truth conditions}} &\sim \text{Uniform}(0,1)\\ \phi_{\text{implied prevalence (by type)}} &\sim \text{Uniform}(0,1)\end{aligned}$$

E.4.2 Posteriors over model parameters

The MAP and 95% Credible interval for the contamination parameter in the implied prevalence task is 0.14 [0.08, 0.18]. The MAP and 95% Credible interval for the contamination parameter in the truth conditions task is 0.43 [0.32, 0.49].

It should be noted that since the contamination parameter ϕ is task-wide parameter, it cannot account for differences observed in the model between property-types. It is possible, however, that the guessing parameter could remove differences that may exist between conditions. This is particularly relevant for the truth conditions task, where we observe no appreciable differences among the average prevalence to assent to the generic (see Figure 5, truth conditions). To rule out this possibility, we ran the models without contamination parameters, and observed the same effects we

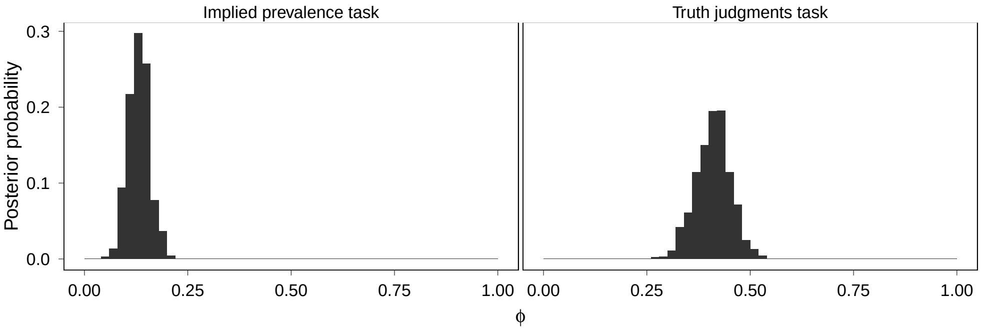

Figure 8: Posterior distribution of the contamination ("guessing") parameter.

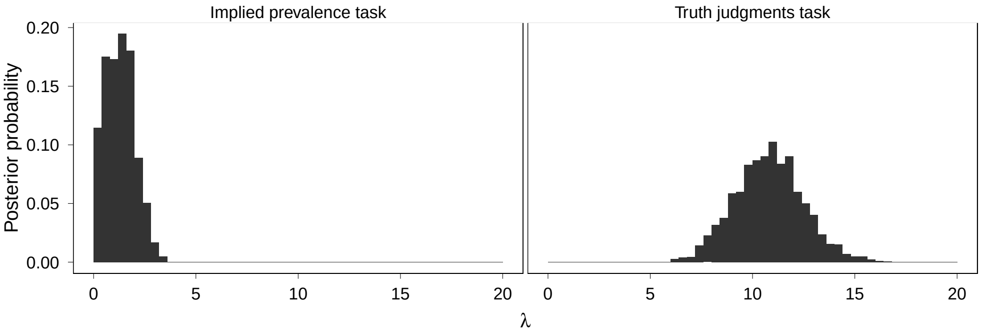

Figure 9: Posterior distribution of the speaker rationality parameters.

report in the main paper: a linear increase as a function of mean prevalence when present for the implied prevalence task, and no linear increase for the truth conditions data. Thus, the insensitivity of the average prevalence at which speakers assent to the generic that we observed in truth conditions task is also a property of our model, and is not a function of the noise process we assume.

The relatively high contamination parameter for the truth conditions task is most likely due to fact that human participants were less willing to assent to the generic *overall* in this task. This has the effect of increasing the average prevalence to assent. To try to account for this quantitatively, the model assumes a relatively high amount of random guessing. We note that a quantitative fit to the average prevalence to assent to novel generics is not of theoretical interest here. Of interest is the relative insensitivity of the average prevalence score to the *a priori* mean prevalence when present, which is a robust prediction of the model.

The estimated posterior distribution of the speaker rationality parameters $\lambda_{\text{truth conditions}}$ and $\lambda_{\text{implied prevalence}}$ are shown in Figure 9. This parameter represents the belief in how rational the hypothetical speaker in Eq. 2 is believed to be when choosing to say the generic (over saying nothing). The MAP and 95% credible interval for $\lambda_{\text{truth conditions}}$ is 11.61 [7.63, 14.3] and $\lambda_{\text{implied prevalence}}$ is 1.54 [0.05, 2.52]. The fact that these parameters are so different is expected given that the response space in the two tasks is also quite different.

E.4.3 Posterior predictives

All model predictions of interest are shown in the main text.

References

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, *34*(8), 1452–1482.

Gelman, S. A., Star, J. R., & Flukes, J. E. (2002). Children's Use of Generics in Inductive Inferences. *Journal of Cognition and Development*, *3*(2), 179–199.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* http://dippl.org. (Accessed: 2015-7-17)

Griffiths, T. L., & Tenenbaum, J. B. (2005, December). Structure and strength in causal induction. *Cognitive psychology*, *51*(4), 334–84. doi: 10.1016/j.cogpsych.2005.05.004

Lee, M. D., & Wagenmakers, E. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge: Cambridge University Press.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, *90*(4), 339–363. doi: 10.1037/0033-295X.90.4.339

Prasada, S., Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2013, March). Conceptual distinctions amongst generics. *Cognition*, *126*(3), 405–22. doi: 10.1016/j.cognition.2012.11.010

Rose, C. E., Martin, S. W., Wannemuehler, K. A., & Plikaytis, B. D. (2006). On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *Journal of biopharmaceutical statistics*, *16*(4), 463–481.

Table 2: Stimuli used in Experiment 1. Estimates are proportion agreement for truth judgments and Maximum A-Posteriori (MAP) estimates for prevalence. Brackets denote 95% confidence intervals for truth judgments and 95% credible intervals for prevalences.

Conceptual type	Item	Truth judgment	Prevalence
Majority characteristic	1. Leopards have spots.	0.956 [0.912, 0.989]	92.7 [85.9, 99.0]
	2. Ducks have wings.	0.945 [0.89, 0.989]	98.5 [95.2, 99.9]
	3. Cardinals are red.	0.989 [0.967, 1]	75.5 [62.1, 86.9]
	4. Swans are white.	0.901 [0.835, 0.967]	67.3 [57.1, 73.8]
	5. Peacocks have beautiful feathers.	0.989 [0.967, 1]	92.0 [77.7, 100]
Minority characteristic	6. Lions have manes.	0.945 [0.89, 0.989]	54.3 [48.1, 63.0]
	7. Kangaroos have pouches.	0.967 [0.923, 1]	69.9 [65.8, 79.6]
	8. Robins lay eggs.	0.934 [0.879, 0.978]	68.5 [61.1, 74.9]
Striking	9. Sharks attack swimmers.	0.879 [0.813. 0.945]	41.8 [32.0, 54.7]
	10. Mosquitos carry malaria.	0.989 [0.967, 1]	47.2 [38.1, 52.9]
	11. Ticks carry Lyme disease.	0.967 [0.923, 1]	42.6 [40.0, 54.2]
	12. Tigers eat people.	0.692 [0.593, 0.78]	37.3 [23.3, 49.9]
False generalization	13. Robins are female.	0.429 [0.33, 0.527]	51.9 [47.8. 55.4]
	14. Lions are male.	0.571 [0.473, 0.67]	49.9 [46.6, 53.9]
	15. Swans are full-grown.	0.725 [0.637, 0.802]	59.7[51.0, 66.0]
	16. Leopards are juvenile.	0.143 [0.077,0.22]	28.3 [22.9, 38.2]
	17. Sharks are white.	0.341 [0.253, 0.44]	32.2 [15.8, 44.6]
False or Uncertain	18. Leopards have wings.	0.022 [0, 0.055]	1.0 [0.0, 4.1]
	19. Kangaroos have spots.	0.033 [0, 0.077]	5.0 [0.1, 13.5]
	20. Tigers have pouches.	0.011 [0, 0.033]	2.0 [0.0, 12.3]
	21. Robins carry malaria.	0.055 [0.011, 0.099]	5.4 [1.6, 9.1]
	22. Sharks have manes.	0.044 [0.011, 0.099]	0.3 [0.0, 5.7]
	23. Lions lay eggs.	0 [0, 0]	0.1 [0.0, 4.3]
	24. Sharks don't attack swimmers.	0.231 [0.154, 0.319]	59.3 [49.8, 75.1]
	25. Ticks don't carry Lyme disease.	0.044 [0.011, 0.088]	55.1 [46.6, 60.8]]
	26. Mosquitos don't carry malaria.	0.077 [0.033, 0.132]	58.7 [50.5, 65.3]]
	27. Tigers don't eat people.	0.297 [0.198, 0.385]	65.3 [56.9, 97.7]
	28. Peacocks don't have beautiful feathers.	0.022 [0, 0.055]	17.4 [0.5, 27.5]
	29. Mosquitos attack swimmers.	0.385 [0.286, 0.495]	26.5 [8.4, 39.2]
	30. Sharks lay eggs.	0.088 [0.033, 0.143]	18.5 [0.6, 41.3]

Table 3: Stimuli used in Experiment 2 and statistics of the priors measure in the prior elicitation task. *Potential to be present* is a measure of how many different kinds are expected to have the property. Mean prevalence when present is a measure of how widespread the property is expected to be, assuming that it is present within a kind. Maximum A-Posteriori (MAP) estimates and 95% Highest Probability Density Intervals from the Bayesian data analysis model described in Section C.3 are shown for each measure. Several of these items are taken from Cimpian et al. (2010)

Property type	Item	Potential to be present θ	Mean prevalence when present γ
Body part	fur	0.78 [0.81, 0.69]	0.90 [0.828, 0.93]
	skin	0.89 [0.93, 0.85]	0.89 [0.853, 0.93]
	feathers	0.68 [0.74, 0.61]	0.90 [0.827, 0.91]
	legs	0.88 [0.93, 0.81]	0.87 [0.819, 0.94]
	tails	0.75 [0.82, 0.65]	0.91 [0.881, 0.93]
	ears	0.85 [0.89, 0.80]	0.88 [0.839, 0.91]
	claws	0.73 [0.74, 0.61]	0.87 [0.749, 0.89]
	teeth	0.83 [0.88, 0.77]	0.88 [0.810, 0.91]
Color	silver legs	0.37 [0.46, 0.30]	0.58 [0.472, 0.64]
	yellow fur	0.53 [0.65, 0.48]	0.69 [0.594, 0.76]
	violet skin	0.39 [0.51, 0.33]	0.63 [0.524, 0.74]
	orange ears	0.46 [0.54, 0.35]	0.60 [0.488, 0.69]
	blue claws	0.38 [0.44, 0.26]	0.66 [0.481, 0.69]
	pink teeth	0.22 [0.37, 0.21]	0.54 [0.421, 0.66]
	orange tails	0.51 [0.57, 0.38]	0.65 [0.581, 0.75]
	purple feathers	0.48 [0.54, 0.36]	0.64 [0.509, 0.72]
Vague	big claws	0.63 [0.70, 0.55]	0.78 [0.682, 0.84]
	long teeth	0.60 [0.66, 0.50]	0.73 [0.694, 0.84]
	rough skin	0.62 [0.72, 0.55]	0.74 [0.672, 0.80]
	curly fur	0.55 [0.64, 0.48]	0.76 [0.669, 0.82]
	long legs	0.63 [0.68, 0.55]	0.76 [0.680, 0.83]
	smooth feathers	0.62 [0.68, 0.50]	0.80 [0.692, 0.83]
	long tails	0.62 [0.69, 0.53]	0.82 [0.738, 0.85]
	small ears	0.65 [0.70, 0.55]	0.81 [0.751, 0.85]
Common accidental	torn tails	0.47 [0.54, 0.32]	0.23 [0.202, 0.35]
	wet fur	0.57 [0.66, 0.47]	0.45 [0.376, 0.56]
	dusty skin	0.45 [0.56, 0.39]	0.44 [0.393, 0.58]
	torn feathers	0.46 [0.58, 0.40]	0.29 [0.218, 0.35]
	fungus-covered fur	0.37 [0.51, 0.29]	0.37 [0.294, 0.49]
	worn-out claws	0.54 [0.62, 0.46]	0.41 [0.336, 0.54]
	muddy feathers	0.41 [0.58, 0.36]	0.39 [0.291, 0.45]
	sore teeth	0.39 [0.51, 0.33]	0.25 [0.188, 0.33]
Rare accidental	broken legs	0.30 [0.44, 0.23]	0.13 [0.098, 0.17]
	swollen ears	0.42 [0.52, 0.32]	0.24 [0.196, 0.34]
	itchy tails	0.37 [0.48, 0.28]	0.33 [0.231, 0.38]
	rotten teeth	0.56 [0.62, 0.42]	0.31 [0.225, 0.39]
	sore legs	0.43 [0.53, 0.34]	0.24 [0.209, 0.34]
	cracked claws	0.50 [0.60, 0.36]	0.24 [0.194, 0.36]
	infected ears	0.42 [0.54, 0.34]	0.18 [0.131, 0.23]
	burned skin	0.32 [0.40, 0.23]	0.19 [0.133, 0.25]