# Words are vague: A prevalence-based account of generic language

**Michael Henry Tessler** (mtessler@stanford.edu)

**Noah D. Goodman** (ngoodman@stanford.edu)

Department of Psychology, Stanford University

April 25, 2015

**Abstract**

Generic utterances (e.g. "Dogs bark") are ubiquitous in natural language. Despite their prevalence, the meanings of generic statements are puzzling to formal approaches. A traditional, scalar semantic account would stipulate a generic statement is true if the prevalence (i.e. the probability of the property given the category) is greater than some threshold. Such accounts are difficult to defend because of the inherent flexibility in generic truth conditions. Here, we propose a scalar semantics for generics wherein the truth conditional threshold is underspecified in the semantics but is determined through pragmatic reasoning. We illustrate the model by showing its ability capture gender-specific and low-prevalence generics—two cases of theoretical importance We use an experimental paradigm developed by Cimpian, Brandone, and Gelman (2010), who first provided empirical evidence for flexibility in truth conditions, as well as a "paradoxical asymmetry" between the truth conditions and implications of generics. In Expts. 1, 3a & 4a, we replicate those findings. Our model predicts flexibility in truth conditions across different types of properties and asymmetry effects *if* the prior distribution over prevalence differs across these property-types. In Expts. 2, 3b & 4b, we find direct evidence that this is so.

**Keywords:** generics; pragmatics; bayesian model

Animals breathe Oxygen to live. (Simple Wikipedia)

Imagine talking with a 3-year-old about oxygen. Oxygen is difficult to explain because it is so rarely observable and the function of oxygen is abstract. You might use a construction similar to the one from Simple Wikipedia above. Few would argue with the truth of this sentence, yet its meaning is difficult to specify.

This type of utterance is generic (Carlson, 1977; Leslie, 2008) in that it conveys a generalization about a category. At first glance, generics statements might seem to mean the same as universally-quantified statements (i.e. "All animals breathe oxygen"). Unlike "all"-statements, however, generics are resilient to counter-examples: "Animals breathe Oxygen to live" is true even though there are three species of creatures called *lorica* that live 2.2 miles underwater on the ocean floor in an environment that is oxygen-free (Danovaro et al., 2010). You might think, then, that generics are synonymous with weaker "most"-statements. However, we'll only be able to maintain this logic until we encounter something like "Birds lay eggs" or "Mosquitos carry West Nile Virus", which are both intuitively true despite the fact that only adult female birds lay eggs and that less than 1% of mosquitos carry West Nile Virus. The only quantifier that could satisfy all of these truth conditions is the existential "some". Yet, the semantics for "some" cannot be the semantics of generics: Upon hearing a generic, listeners are wont to interpret the sentence as applying to *all or nearly all* of a category (S. A. Gelman, Star, & Flukes, 2002; Cimpian et al., 2010). "Some animals breathe oxygen" implies something quite different (Degen, 2015).

Awesome paragraph

What could be the stable meaning of a generic given this extreme flexibility? We propose these phenomena can be explained as the effects of pragmatic inference filling in a meaning that is underspecified in the semantics. In particular, we posit a scalar semantics for generics in which they express that the probability of the property given the kind——i.e.

the property's *prevalence*——is above a threshold (cf. Cohen (1999)). Following Goodman and Lassiter (2015), we treat this threshold as underspecified in the semantics and thus, as a free variable that is reasoned about by a pragmatic listener within the *Rational Speech-Act* (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013). In this way, generics are vague in the way gradable adjectives like "tall" are vague. In this formalism, the listener is tasked with the joint problem of figuring out what the speaker intended to communicate in addition to the what the threshold is likely to be (i.e., what does the generic mean *in this situation*?). With an uncertain semantics, the listener solves this puzzle by drawing upon her prior beliefs, in this case about the prevalence of the property, in addition to harnessing standard inferences from conversational pragmatics (Clark, 1996; Grice, 1975; Levinson, 2000). From this, flexible truth conditions can be derived. At the same time, generics will have near-universal interpretations if the prevalence of the property follows particular distributions.

Cimpian et al. (2010) (henceforth, CBG) carried out a series of experiments designed to examine the truth conditions and implications of generic statements about novel animal categories (e.g. "lorches have purple feathers") . They found empirical support for the flexibility of truth conditions by manipulating the type of property in question (e.g. if the property was *distinctive* or *dangerous*); this manipulation led to differences in the acceptably of a generic at low levels of prevalence (e.g. when only "30% of lorches have purple feathers"). CBG also reported an asymmetry between the truth conditions entailed by this *verification* task and those reported in an analogous *interpretation* task, where participants were asked to provide a prevalence estimate upon hearing a generic. The title of their article provides an excellent summary: "Generic statements require little evidence for acceptance but have powerful implications"[1].

In what follows, we describe our formal proposal for generic meaning and test predictions in a number of experimental conditions. We demonstrate that the model captures cases of theoretical importance: gender-specific and low-prevalence generics. We replicate the main effects reported by CBG (Expt. 1) and use Bayesian data analytic techniques to further how the effective truth-conditions of generic statements interact with the prior on prevalence. We show that our model predicts both flexible truth conditions and asymmetry effects, given appropriate priors over prevalence ("*backward predictions* of the model"). We experimentally elicit the prevalence priors in CBG's experimental contexts, verifying these predictions (Expt. 2). We show that the elicited prevalence priors predict the same effects as the inferred priors ("*forward predictions* of the model"), and then show how the effects dissipate when different types of properties (with different types of priors) are discussed (Expt. 3). Finally, we explore a case when the flexibility of truth conditions are difficult to associate with prevalence (Expt. 4).

# 1 A lifted-threshold model of generic meaning

We view language understanding as a special case of social cognition, and language comprehension as deriving from an intuitive theory of language production. We draw on recent work from probabilistic pragmatics to formalize how listeners arrive at interpretations of generic utterances. In particular, we draw on work from the Rational Speech-Act (RSA) theory of language understanding. In this framework, a listener infers the meaning of an utterance by considering the thought-processes of a speaker whose goal is to be informative. Variants of this theory have provided formal explanations for a number of linguistic phenomena including scalar implicature, hyperbole, and argument evaluation (Kao, Wu, Bergen, & Goodman, 2014; Tessler & Goodman, 2014; Lassiter & Goodman, to appear).

We propose that generics are vague. In this way, generics are similar to gradable adjectives like *tall*. For example, the conditions under which "John is tall" and "The Empire State Building is tall" are quite different. Lassiter and Goodman (2015) propose the meaning of an adjective like *tall* is a standard truth-functional, threshold meaning such that the object in question *is tall* if it has a height greater than the threshold $\theta_{tall}$. The vagueness and context-sensitivity

[1] These basics findings were also replicated in young children(Brandone, Gelman, & Hedglen, 2014)

of these adjectives are accounted for by treating $\theta_{tall}$ as an unknown property of the language, and modeling the pragmatic listener as inferring this threshold. In this example, different thresholds are derived through the interaction of the prior distribution over heights (people vs. buildings) with the communicative pressures to be truthful and informative.

In a similar way, we propose the literal semantics of a generic sentence is in fact a threshold on prevalence, but listeners have uncertainty about the threshold *a priori* and actively must reason about it in context. The relevant prior distribution is the distribution of the property (e.g. "breathes oxygen" will have a different distribution than "breathes polluted air"). This is the distribution of the property prevalence across categories (e.g. 0% of books breathe oxygen, 10% of Americans breathe polluted air, 100% of dogs breathe oxygen). Just as the prior distributions of heights for people and buildings differ, the prior distributions over prevalence for different types of properties can vary.

The RSA model for generic interpretation, with the prevalence threshold as a variable "lifted" to pragmatic reasoning is specified by:

$$P_{L_0}(x \mid g, \theta) \propto \delta_{[\![g(\theta)]\!](x)} P(x) \tag{1}$$

$$P_{S_1}(g \mid x, \theta) \propto \exp(\lambda \ln P_{L_0}(x \mid g, \theta)) \tag{2}$$

$$P_{L_1}(x, \theta \mid g) \propto P_{S_1}(g \mid x, \theta) P(x) \tag{3}$$

Here $[\![g(\theta)]\!] : X \rightarrow$ Boolean is a truth-function specifying the literal meaning of the generic. The literal content in Eq. (1) is given by $[\![g(\theta)]\!] = \{x | x > \theta\}$, where $x$ is a prevalence. $g$ is a function of $\theta$ because the meaning of a generic may vary across contexts.

Eq. (3) is a model of a listener ($L_1$) who has been told a generic statement. She assumes that, whatever prevalence $x$ the speaker ($S_1$) meant to communicate, the speaker was trying to be informative. By using a vague utterance and mentioning a particular property (e.g. "breathes oxygen"), the speaker is, in essence, also communicating the actual meaning of the generic $\theta$ in this situation. The listener assumes the speaker in Eq. (2) knows $\theta$ and chooses an utterance to be informative to a hypothetical literal listener ($L_0$). The degree to which the speaker's utterance is optimal for a given intended-prevalence is governed by a soft-max decision rule with rationality parameter $\lambda$ (Luce, 1959). From this, the pragmatic listener ($L_1$) jointly infers both the prevalence $x$ and the threshold $\theta$. We call this type of model a "lifted threshold" model because $\theta$, traditionally thought to be part of the semantic content of the utterance (and thus perfectly transparent to all in the conversation), has been underspecified in the semantics but is locally fixed through pragmatic reasoning.

## 1.1 Simulations of theoretical interest

We propose that a model of generics with a scalar semantics on prevalence is tenable if the threshold is left underspecified in the semantics. This section explores the influence of different prevalence priors on different truth conditions for a few examples of theoretical interest. To connect our model of generic interpretation to generic truth conditions, we include an additional component to the model: a speaker who can either say "[[the generic]] is true" or "[[the generic]] is false" (Degen & Goodman, 2014):

$$P_{S_2}(g \mid x) \propto \exp(\lambda \ln P_{L_1}(x \mid g)). \tag{4}$$

This speaker in (4), like $L_1$, doesn't know the threshold, but knows that $L_1$ is thinking about it, and marginalizes over possible values: $P_{L_1}(x \mid g) = \sum_\theta P_{L_1}(x, \theta \mid g)$. The $S_2$ speaker has in mind a prevalence $x$ (or, equivalently here, a category–property pair corresponding to that prevalence), e.g. "birds" and those that "lay eggs". The speaker then reasons whether it would be better to say "'Birds lay eggs' is true" or "'Birds lay eggs' is false" in order to convey the

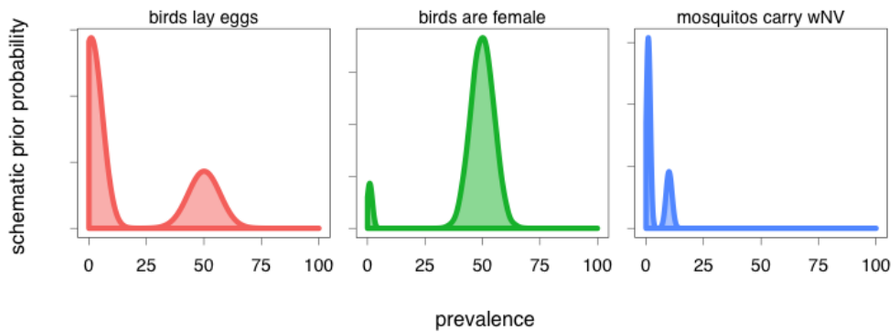



Figure 1: Schematic priors over prevalence for generics like “Birds lay eggs”, “Birds are female”, and “Mosquitos carry West Nile Virus”.

state of the world, namely that about 50% of birds lay eggs[2]. The speaker selects the utterance according to a soft-max decision rule governed by parameter $\lambda$[3].

Continuing with this example, the speaker imagines the likely prevalence that the pragmatic listener $L_1$ will infer given either “‘Birds lay eggs’ is true” or “‘Birds lay eggs’ is false”. Here, the prevalence prior plays a critical role in determining where the generic-threshold $\theta$ (and, consequently, the resulting inferred prevalence $x$) is likely to fall (Figure 1). In setting the threshold, the pragmatic listener $L_1$ balances the truth of the utterance with the informativeness of the utterance. If “‘Birds lay eggs’ is true”, the pragmatic listener $L_1$ will likely set the threshold $\theta$ below 50%, as this is necessary to make the utterance true (Figure 1; red). At the same time, $L_1$ believes the message to be informative, so she will set $\theta$ probably above 10%. Since the prevalence has to be above the threshold in order to make the utterance true, the most likely inferred prevalence $x$ will be about 50% (i.e. after ruling out everything below 10%, the most probably prevalence will be 50%). On the other hand, if the speaker were to say “‘Birds lay eggs’ is false”, the prevalence would have to fall *below* the threshold, which in this case would implicate the mass below 10% and the pragmatic listener would infer something the speaker didn’t intend.

A natural foil for this example is “Birds are female”. For birds, “lay eggs” seems to have the same prevalence as “are female” but many speakers judge the generic “Birds are female” to be false (Khemlani, Leslie, & Glucksberg, 2009; Brandone, Cimpian, Leslie, & Gelman, 2012). “Birds are female” is different from “Birds lay eggs” because the overwhelming majority of animal kinds are about 50% female (see Figure 1; green). By contrast, very few animal kinds have any egg layers; those animal kinds that do have about 50% egg layers (Figure 1; red). The lifted threshold RSA model is ambivalent between the saying “‘Birds are female’ is true” and “‘Birds are female’ is false”, while at the same time, has a clear preference for supporting “Birds lay eggs” (Figure 2; medium prevalence, red vs. green bars).

Another example of theoretical concern is a bimodal prior with a peak at some low prevalence level. This prior should describe properties that are not only rare *across kinds* but also rare *within kinds* (Figure 1, blue). A canonical example of this is “West Nile Virus” in the generic “Mosquitos carry West Nile Virus”. For a prior like this, the lifted-threshold RSA model predicts the truth conditions would be relaxed at low prevalence levels (Figure 2; low prevalence).

## 2 Experiment 1: CBG – primary findings replication

The lifted-threshold RSA model predicts different truth conditions for generics when the prevalence prior differs. A paradigm was developed by Cimpian et al. (2010) to get at a similar question. CBG also was concerned with

[2]Technically, the prevalence of birds laying eggs is closer to 35%, as only *adult*, female birds lay eggs. We gloss over this detail here.

[3]It’s conceivable that this $\lambda$ is different from the parameter that governs the selection of $S_1$’s utterances. For simplicity here we assume they are the same.

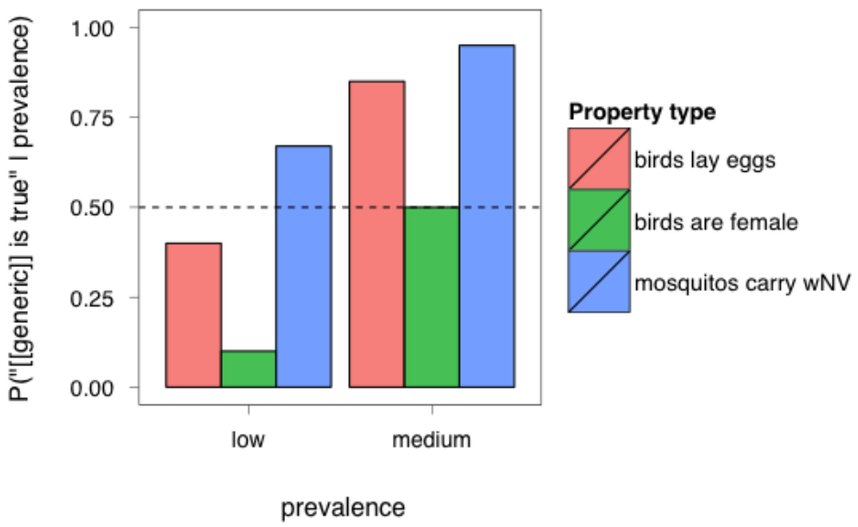



Figure 2: Schematic truth conditions for generics like "Birds lay eggs", "Birds are female", and "Mosquitos carry West Nile Virus". 0.5 denotes the point at which the generic is equally true and false.

determining whether or not the truth conditions for and implications from generic statements were comparable. To flesh out the motivation for the second question: consider sentences with the quantifier "All" (e.g. "All lorches have purple feathers"). Presumably, this sentence is true only when 100% of lorches have purple feathers. Similarly, upon hearing such an utterance, one is likely to infer that 100% of lorches have purple feathers. CBG tested this symmetry with the quantifier "most"— "the quantifier that comes closest to capturing generic meaning (Carlson, 1977; Cimpian & Cadena, 2010)". They found that participants judged "most" sentences true when between 50-100% of the category had the property. Similarly, upon hearing an utterance with "most", participants on average inferred a prevalence of about 75%. Thus, "most" also retains this symmetry as well. CBG found that generic statements, however, are judged true for a wide range of prevalence levels, but upon hearing a generic utterance, participants were likely to infer that *all* or *almost all* of the category had the property. Thus, the generic shows an *asymmetry* between truth conditions and implications.

Experiment 1 attempted to replicate the main findings of CBG: that the type of property affects the proportion of "true" responses to a generic statement (Exp. 1a) and that there is an asymmetry between truth conditions and implications of the generic (Exp. 1b).

## 2.1 Experiment 1a: *truth conditions*

In CBG's *truth conditions* task, participants are given an evidence statement consisting of the percentage of a novel animal category that had a property (e.g. "30% of lorches have purple feathers"). Participants were asked to judge the associated generic statement (i.e. "Lorches have purple feathers") as true or false.

Following CBG, we manipulated both the prevalence and the type of property within-subjects. Prevalence varied between 10, 30, 50, 70, and 90%. Property type was manipulated by adding additional sentences to the prompt. CBG's original study used three property types: *dangerous and distinct* (e.g. "These feathers are as sharp as needles and can easily get lodged in you, causing massive bleeding. No other animals have these kinds of feathers."), *nondangerous and nondistinctive* (e.g. "These feathers are wide and very smooth to the touch. Other animals have these kinds of feathers."), and *plain* (no additional statements).

CBG found that proportion of "true" responses increased monotonically as prevalence increased. They also found an interaction with type: *dangerous and distinctive* property had higher proportions of "true" responses to the generic, particularly so at lower prevalence levels.

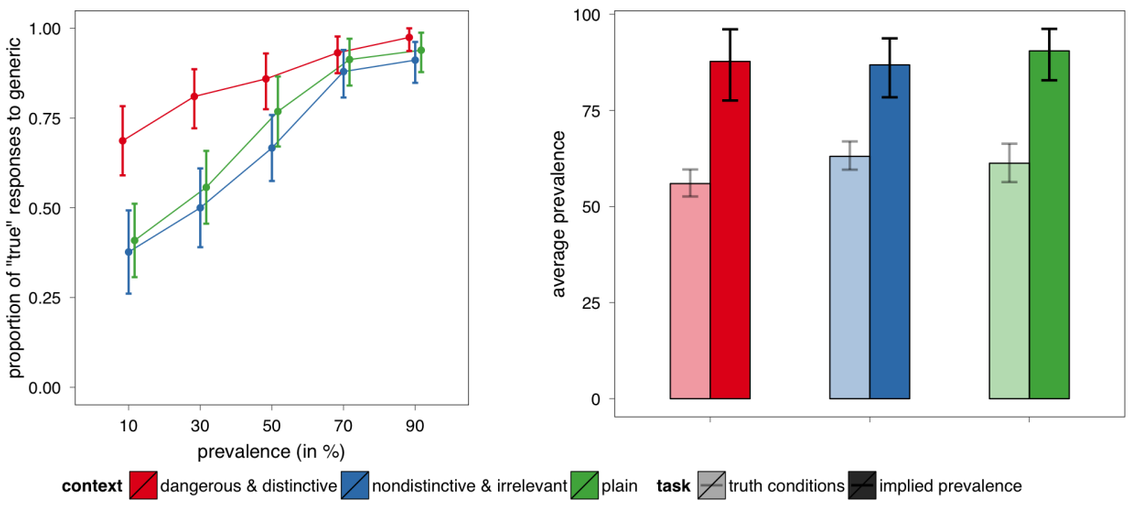



Figure 3: Replication of CBG. Left: *truth conditions* vary by context (Exp. 1a). Right: *implied prevalence* of the generic is greater than *truth conditions* (Exp. 1b). Error bars denote bootstrapped 95% confidence intervals.

### 2.1.1 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 $0.50.

### 2.1.2 Procedure and materials

Our procedure was very similar to CBG's *truth conditions* task. Participants were told they were the resident zoologist of a team of scientists that recently discovered an island with many new animals; their task was to provide their expert opinion on questions about these animals[4].

We used the same materials as CBG (available in their Appendix). The materials used were 30 novel animal categories (e.g. lorches, morseths, blins) each paired with a unique property. Properties were made by pairing a color with a body-part (e.g. purple feathers, orange tails). Each participant saw 30 unique animal-property pairs: 10 of each of the 3 types (*dangerous and distinct*, *nondangerous and nondistinct*, *plain*). The 10 items of each property-type were randomly paired with 1 of 5 "prevalence levels": $\{10, 30, 50, 70, 90\}$%; thus, each prevalence level appeared 2 times per type.

On each trial, participants saw a prevalence statement and type statements (*dangerous and distinct*, *nondangerous and nondistinct*, *Plain*; illustrated above). Participants were then asked "Is the following sentence true or false?", below which was presented the associated generic (e.g. "Lorches have purple feathers") and "True" and "False" radio buttons.

### 2.1.3 Results

Results are shown in Figure 3 (Left). We entered participants' truth judgments into a mixed effects logistic regression with random by-item and by-participant effects of intercept and fixed effects of prevalence and type as well as their interaction[5]. Our results replicated the finding of CBG that the generic statements were endorsed more with dangerous and distinctive properties than with plain properties (Figure 3, left; $\beta = 1.99; SE = .36; z = 5.52; p < .001$). There was also an interaction between prevalence level and type such that the generic was endorsed more with *dangerous and distinctive* properties than with plain properties at lower prevalence levels ($\beta = .03; SE = .01; z = 2.35; p = 0.019$). There was a trending effect for the *nondangerous and nondistinct* properties to be endorsed *less* than the plain properties

[4]The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/cbg2010-replication/experiment/experiment-9.html`

[5]This was the maximal mixed-effect structure supported by the data.

($\beta = -.57; SE = .30; z = -1.91; p = .056$).

## 2.2 Experiment 1b: *implied prevalence*

In their *implied prevalence* task, participants were supplied with the generic and asked to judge prevalence: "What percentage of lorches do you think have purple feathers?". Type was again manipulated within-subject. CBG found that the generic was interpreted strongly—nearly all lorches have purple feathers—for all three types of properties.

### 2.2.1 Participants

We recruited 30 participants over MTurk[6]. 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 $0.50.

### 2.2.2 Procedure and materials

Our procedure was very similar to CBG's *implied prevalence* task. Our instructions were the same as in Exp. 1a[7].

The materials and property-type conditions were the same as in Exp. 1a. Each participant saw 10 trials of each of the 3 property-types (30 trials in total). On each trial, participants saw a generic statement and property-type statements. Participants were then asked "What percentage of [the kind] do you think have [the property]?" (e.g. "What percentage of lorches do you think have purple feathers?") The dependent measure was a free response required to be an integer, $0-100$.

### 2.2.3 Data analysis and results

To compare the truth conditions data with the implied prevalence data, we followed the data analysis strategy of CBG. Using the data from Exp. 1a, we computed, for each subject, an *average prevalence level* that led to "True" responses. For example, if a participant said "True" whenever the prevalence was 70% or 90% and "False" to everything else, that participant received an *average prevalence score* of 80%; if a participant said "False" to everything, their *average prevalence score* was 100%, since they presumably were interpreting the generic statement as a universal (i.e. an "all" statement). This score was compared against the implied prevalence dependent measure of Exp. 1b.

The prevalence scores from each task were entered into a linear mixed model with a by-participant random effect of intercept; the fixed effects were property-type, task, and their interaction. Our results replicated the asymmetry finding of CBG that the generic statement was interpreted as having a higher prevalence than its truth conditions entail (i.e. main effect of task; $\beta = 28.8; SE = 4.3; t = 6.6; p < 0.001$; see Fig. 3, right). In the original study, this asymmetry was not observed for sentences using the quantifier "most"; however, we did not replicate this effect here.

# 3 Model analysis

A schematic of the model and our linking assumptions is shown in Figure 4. We model the *truth conditions* task as a speaker who can either say "[[the generic]] is true" or "[[the generic]] is false", as we did in Section 1.1. This speaker could also be thought of as having a second $L_1$ branching off from her, charged with interpreting the prevalence statement given in the task (e.g. "30% of lorches have purple feathers"). This submodel would reduce to a delta function, as the utterance ("30% ...") is completely unambiguous and maps directly onto a prevalence level (i.e. "30%").

[6] This study was ran 1 week after Exp. 1a. None of the participants in Exp. 1a participated in Exp. 1b.

[7] The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/cbg2010-replication/experiment/experiment-12.html`

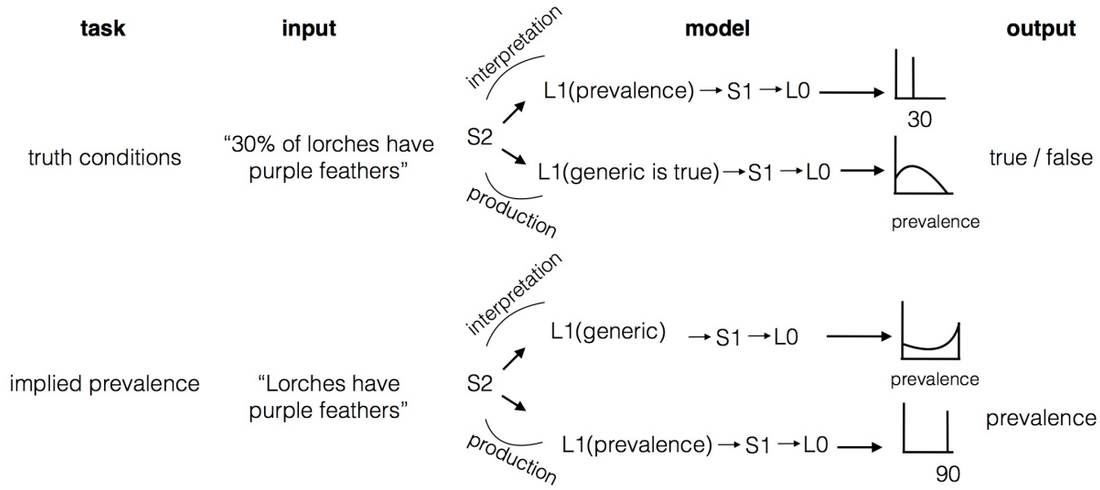



Figure 4: Schematic of the input–output structure of the model.

We model the *implied prevalence* task also as an $S_2$. Just like the participant in the task, this model takes the generic as input. The generic then gets passed immediately down to the $L_1$ listener, who does lifted-threshold inference to determine the right interpretation. The *implied prevalence* $S_2$ then says which prevalence is most likely to be the case. This "prevalence utterance" gets passed down to an $L_1$. Like the prevalence $L_1$ in the *truth conditions* model, this submodel reduces to a delta function (i.e. responding "90% of lorches have purple feathers" means that 90% of lorches have purple feathers)[8]. We articulate both models as $S_2$s to highlight the similarities between them.

## 3.1 Bayesian model evaluation

As a first test of our lifted-threshold model of generics, we posit a family of possible priors over the prevalence $x \sim \beta(\gamma, \delta)$[9]. We hypothesize that the details of these priors (i.e. $\gamma$'s and $\delta$'s) may differ according to the type of property in reference by the generic. For instance, when you know that a particular property is rare, a different prior distribution of that property over kinds (i.e. a different prevalence prior) is called to mind, than if the property is common. This would result in different meanings for the generic. Below we infer appropriate prior parameters for each property-type from the behavioral data.

We infer the parameters of the prevalence prior, $\beta(\gamma, \delta)$ using uninformative hyperpriors:

$$\gamma_{type} \sim U(0,1)$$
$$\delta_{type} \sim U(0,5)$$
$$\phi_{task} \sim U(0,1)$$
$$\lambda_{task} \sim U(0,5)$$

where $type \in \{$dangerous and distinct, nondangerous and nondistinct, plain$\}$ and $task \in \{$truth conditions, implied prevalence$\}$.

We account for inattention and other irrelevant factors by including a probability $\phi_{task} \sim U(0,1)$ for each task that a given response is the result of uniform random guessing[10] (Lee & Wagenmakers, 2014). The inferred "guessing" parameter $\phi$ is the amount of data that would have to be attributed to random guessing in order for our cognitive model

[8] N.B.: This model as a whole is equivalent to just the pragmatic listener ($L_1$) model that is trying to infer the prevalence.

[9] For ease of interpretation, we are parametrizing the $\beta$ distribution by its mean and concentration. To recover the canonical shape parametrization, use $\gamma\delta$ and $(1-\gamma)\delta$.

[10] Ideally, we would have $\phi$ be a function of participant (some participants guess more than others) and experimental condition (some conditions are more difficult or less constrained and invite more guessing). This is computationally too demanding when coupled with the complex cognitive model explored here.

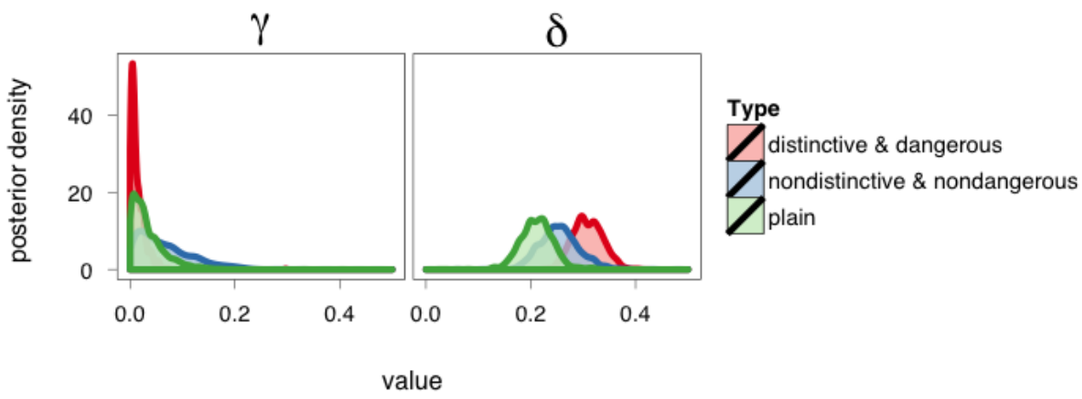



Figure 5: Posterior distributions of the hyperprior parameters used in the lifted-threshold generics model.

of the generic to apply to the experimental data. In this sense, $\phi$ gives a coarse notion of model fit.

### 3.1.1 Inferred parameters

The mean inferred values of $\phi_{1a}$ and $\phi_{1b}$ are about 0.08 and 0.05, respectively, a reasonable rate of "guessing" for participants on Amazon's Mechanical Turk. This also indicates that our model of cognition is doing a good job at accounting for the signal in participants' responses (i.e., it's better than a model of random guessing). The Maximum A Posteriori (MAP) inferred values for the rationality parameters $\lambda_{truth}$ and $\lambda_{implied}$ were approximately 1 and 2, respectively.

Figure 5 shows the posterior distributions of the hyperprior parameters, $\gamma$ and $\delta$, for the lifted-threshold RSA model. The posterior means for the hyperparameters $\gamma_{type}$ are well-ordered: dangerous & distinct $<$ plain $<$ nondangerous and nondistinct. $\gamma_{type}$ reflects the mean of the prevalence prior. Thus, this can be directly interpreted as the mean prior prevalence for the three property types: *dangerous and distinctive* properties are more rare than the other two types of properties.

Additionally, the $\delta$'s are much lower than 1 in each case, indicating bi-modal priors peaked at 0 and 1. This is consistent with all of the properties being construed as *biological* properties. Biological properties have the feature of being almost universally present or almost univeraslly absent within a kind (e.g. 100% of birds have wings; 0% of humans have wings). The posterior means for the $\delta$'s are also ordered, suggesting that participants may treat variance of prevalence as higher for the *dangerous and distinctive* properties (or simply be more confused).

An intuitive way to visualize these inferred hyperparameters is to marginalize over the posterior parameter values to reconstruct a "canonical" prior distribution over prevalence for each property type. Figure 7a shows these prior distributions inferred from Exp. 1a & 1b data via the lifted-threshold RSA model. Qualitatively, they are each bi-modal and the *dangerous and distinctive* prior has a lower mean.

### 3.1.2 Posterior predictives

We further 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 cognitive model and the observed data. This is akin to fitting the parameters and is an important step in model validation as it shows what data is actually predicted by the model.

The posterior predictions by the lifted-threshold RSA model for the *truth conditions* task are shown in Figure 6. We can see that the model predicts monotonically increasing endorsement rates for the generic as a function of prevalence. The model has some persistent uncertainty about the true value of the threshold and this uncertainty is evident in the human behavior curves of Exp. 1a. The model also matches the differences in endorsement rates between property-type conditions: dangerous and distinctive properties are endorsed more than the other two property types, and this

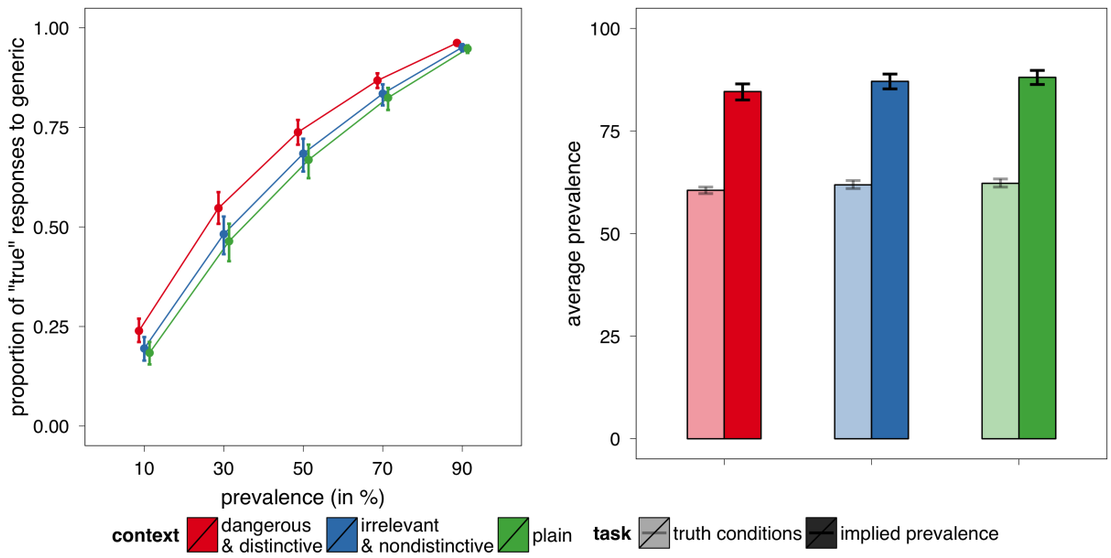



Figure 6: Posterior predictives of lifted-threshold RSA for truth conditions (left) and asymmetry between truth conditions and implications (right).

difference dissipates at high prevalence levels. We reconstruct the curves of Figure 3 reasonably well; the model–data correlation is $r = 0.90$.

We use a similar data analysis strategy as we did for Exp. 1b to compare "average prevalence" between truth conditions and implications. For the *truth conditions* task, we used the model's posterior probability of saying "true" at each prevalence level to simulate trials of the experiment as Bernoulli trials. We simulated 30 trials for each of 1000 imaginary subjects in this way. We then followed CBG's data analysis strategy (as recapitulated in Exp. 1b). The model gives a posterior distribution over prevalences, whose expectation we used to model the *implied prevalence* task. We find the model predicts the asymmetry between interpretation and verification of the generic for all three property types (see Figure 6, right).

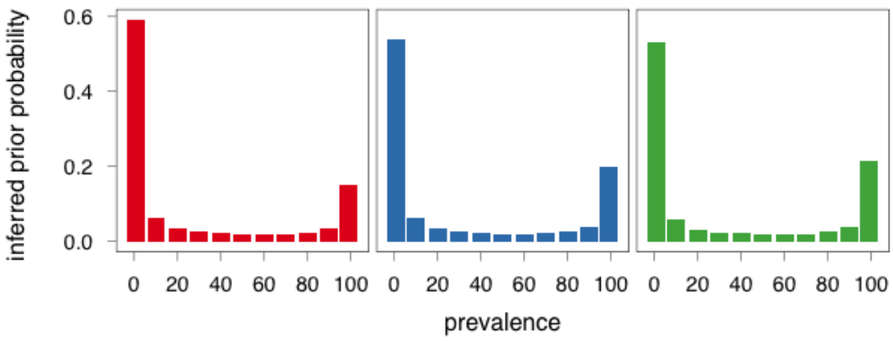
### 3.2 Discussion

To see how this asymmetry is possible, consider again the inferred prevalence priors in Figure 7a. They are bimodal with peaks around 0% and 100%. This is consistent with the intuition that biological properties, such as the ones used by CBG, are properties either held by all of a category or none of a category. Since the semantics of the generic is underspecified (i.e. $\theta$—the threshold for truth judgement—is unknown), if $\theta$ falls anywhere in the range between 10%-90%, the most likely prevalence is going to be near 100% (i.e. after ruling out 0%, the next most-likely alternative is 100%). Hence, in the *implied prevalence task*, the most likely inferred prevalence could be appreciably higher than one would expect from the *truth conditions* task.

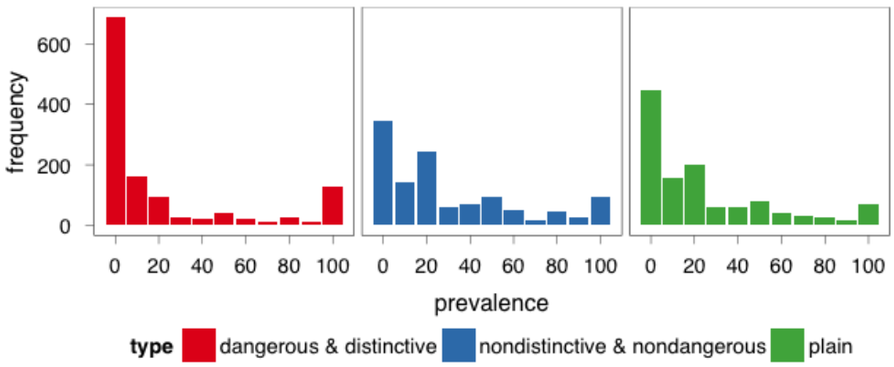
Often in Bayesian data analysis, the posterior distribution over parameters is hard to interpret in terms of observable phenomena. Our case is not so opaque: if our model is the correct model of this task, the prior distributions of prevalence for the three property types should look like they do in Figure 7a. In particular, all three types of properties should have bimodal prior prevalence distributions, with a high probability that 0% of the kind have the property. Further, this left skew should be more pronounced for the *dangerous and distinct* properties relative to the *plain* properties.

## 4 Experiment 2

Exp. 2 sought to test the prediction that the prior distribution of prevalence levels would be bimodal and vary by property type. The proper way to elicit prior beliefs about domains remains an open question in cognitive science.




(a) Reconstructed priors from marginalized posterior $\gamma$ and $\delta$, for each property type.




(b) Priors elicited in Experiment 2.

Figure 7: Prior distributions over prevalence.

Here, we try 4 different dependent measures and assess their reliability with respect to each other.

## 4.1 Method

### 4.1.1 Participants

We recruited 120 participants over Amazon's crowd-sourcing platform Mechanical Turk. 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 4 minutes and participants were compensated $0.40.

### 4.1.2 Procedure and materials

Our procedure[11] was similar to Exp. 1b. On each trial, participants either read information about the property type (*dangerous and distinctive* or *nondangerous and nondistinctive*) or nothing (*plain*).

In addition to the contextual information, participants were told: "Listed below are X kinds of animals that are found on the island." and asked the following question: "What percentage of each kind of animal do you think has [property]?"

Participants were either presented with 1, 5, or 10 animal-names per trials (between-subjects; "Listed below is/are $\{1, 5, 10\}$ kind(s) of animal(s) ..."). For these three groups (n=30 for each), the dependent measure was a free response (or 5 or 10) restricted to be a number between 0–100. A fourth group of participants (n=30) was run in just the 5 kinds/trial condition but with a dependent measure that was a slider bar that ranged from 0–100. The motivation for including multiple kinds per trial was to encourage participants to think about the *distribution* of the property across animal kinds.

In total, participants gave 30 responses (equal numbers for each property type).

[11]The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/cbg2010-replication/experiment/experiment-11.html`

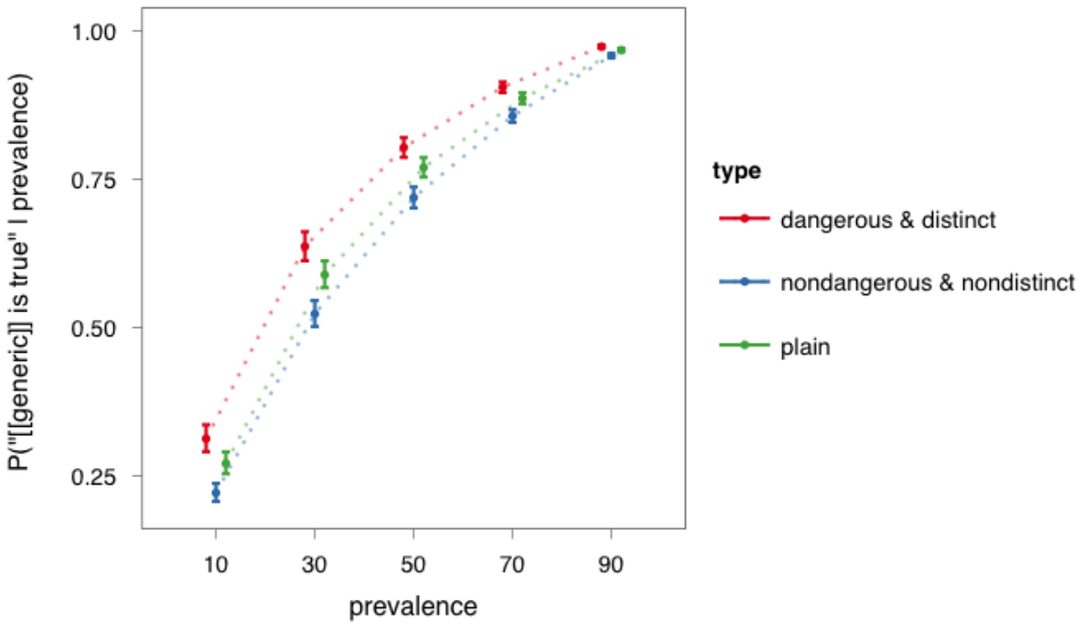



Figure 8: "Forward" predictions of lifted-threshold RSA for truth conditions using the empirical priors from Expt. 2.

## 4.2 Results

The different dependent measures explored produced highly reliably results. QQ-plots revealed a strong linear relationship between the distributions of responses for each of 4 dependent measures used (95% CI for average $r_{pearson} = [0.90, 0.96]; r_{spearman} = [0.93, 0.97]$). Hence, we collapsed across these dependent measures.

Experiment 2 recovered the shape of the inferred prior distributions predicted from the Bayesian analysis of the lifted-threshold RSA model (compare Figure 7b to Figure 7a). Hartigans' Dip Test for Unimodality was highly significant for each of the prior distributions ($D = 0.054, 0.084, 0.0745$ for types *dangerous and distinctive*, *nondangerous and nondistinctive*, and *plain*, respectively; $p < 0.0001$ for each), and thus the distributions are at least bimodal. The means of these three distributions are distinct and ordered as predicted (bootstrapped 95% confidence intervals in parentheses): $\mu_{dangerousdistinctive} = 18.1\%(16.0, 20.2), \mu_{plain} = 20.8\%(19.0, 22.5), \mu_{nondangerousnondistinctive} = 25.7\%(23.7, 27.6)$. The medians of these three distributions were all significantly different from one another, evidenced by pair-wise Mann-Whitney U tests (*dangerous and distinctive* vs. *plain*: $W = 417452$; *dangerous and distinctive* vs. *nondangerous and nondistinctive*: $W = 376180.5$; *nondangerous and nondistinctive* vs *plain*: $W = 548994.5$; all $p < 0.00001$). Finally, the distributions themselves were all significantly different from one another, by Kolmogorov-Smirnov tests (*dangerous and distinctive* vs. *plain*: $D = 0.185$; *dangerous and distinctive* vs. *nondangerous and nondistinctive*: $D = 0.253$; *nondangerous and nondistinctive* vs *plain*: $D = 0.091$; all $p < 0.001$). In sum, the elicited prior distributions are all at least bimodal, have different central tendencies, and are all distinct in shape.

## 4.3 Extension: Using the empirical priors

In Section 3.1, we posited a family of priors for our model (the $\beta$ family of distributions). We saw that our model could accommodate both the flexibility in truth conditions as well as the asymmetry between truth conditions and implications. This combined data-analysis — cognitive model made the *backward prediction* that the priors would have to vary between property-types. In Exp. 2 we found that to be case [12]. There is still a question of whether or not the priors elicited in Exp. 2 would actually still predict the flexible truth conditions and the asymmetry.

To test this, we fix the rationality parameters $\lambda_{task}$ as the fit-value to be the Maximum A-Posteriori (MAP) values

[12] Qualitative differences between the *observed* and *backward-predicted* priors are likely a results of the family of priors posited a priori. The $\beta$ family can only accommodate U- and N- shaped distributions.

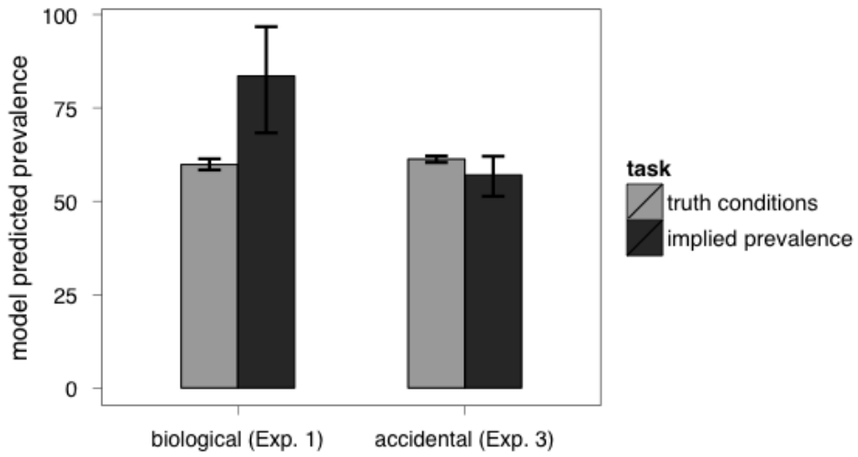



Figure 9: Lifted-threshold RSA coupled with empirical priors predicts strong implications for the "biological" properties used in Expt. 1, but not for the "accidental" properties used in Expt. 3.

from the Bayesian model analysis above. We then sample, with replacement, subjects from our prior elicitation task in order to bootstrap our model predictions. The confidence intervals thus reflect the uncertainty in our model predictions attributable to uncertainty in the empirical prior data[13].

Figure 8 shows the model predictions for the truth conditions using the empirical prior data from Expt. 2. Like the model with β family hyperpriors, the empirical prior model captures the flexibility in truth conditions across the three property types. Figure 9 shows the predicted "strong implications" for the biological type properties used in Expt. 1.

We have seen thus far how a model that takes into account not only the prevalence of a particular property *within* a category but also crucially *across categories* can explain the flexibility in truth conditions of a number of different types of properties. We have also seen how bimodal "biological" priors can lead to near-universal implications for generic statements. In the experiments that follow, we explore cases where these effects break down.

# 5 Experiment 3: Accidental properties

The lifted-threshold model of generic meaning makes the further prediction that should the prior distribution over prevalence of the property *not* have a peak at 100%, the asymmetry should go away. That is, if the prior doesn't have a peak at 100%, the *implications* of a generic statement would not be as strong. CBG also made a similar prediction, though from a different theoretical perspective. They ran a version of the task using *accidental* or *temporary* states and found that the responses in the *implied prevalence* task were significantly reduced (relative to the *biological* type properties used in Exp. 1). In Experiment 3a, we sought to replicate this finding. In Experiment 3b, we measured the prior distribution over prevalence for these *accidental* or *temporary* properties.

## 5.1 Experiment 3a: truth conditions and implications

Experiment 3a differed from Exp. 1a &b only in that there was only one property type: accidental or temporary properties.

### 5.1.1 Participants

We recruited 100 participants over Amazon's crowd-sourcing platform Mechanical Turk. 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 $0.50.

[13]The model predictions are generated using exact enumeration, so there uncertainty in the model predictions due to posterior sampling error.

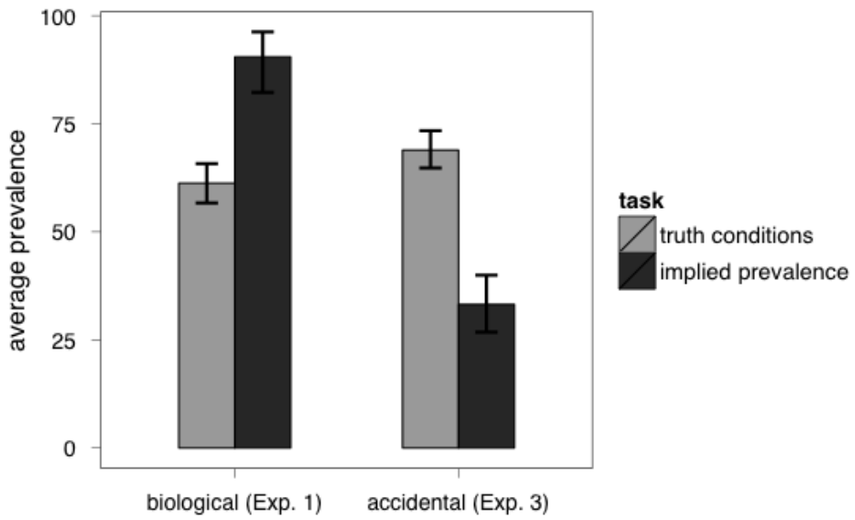



Figure 10: Implied prevalence of generic statements about biological (Exp. 1) and accidental (Exp. 3) are differentially associated with the truth conditions. Error bars denote bootstrapped 95% confidence intervals.

#### 5.1.2 Procedure and materials

Participants were randomly assigned to either the *truth conditions* task (n=59) or the *implied prevalence* task (n=41). Each task consisted of 20 items, all of which referred to temporary, accidental, or disease states (see CBG Appendix B for full list). In the truth conditions task, participants were given a prevalence statement (e.g. "30% of lorches have muddy feathers") and asked if the corresponding generic statement (i.e. "Lorches have muddy feathers") was true or false. In the implied prevalence task, participants were given a given statement (e.g. "Lorches have muddy feathers") and asked to estimate the percentage of the kind that displayed that feature (i.e. what percentage of lorches have muddy feathers?)

#### 5.1.3 Results

We followed the data same data analysis strategy as in Exp. 1 and as in CBG. We replicate CBG's finding that the generic statement about accidental properties was not interpreted as applying to nearly all of the category. Figure 10 shows that difference between the two tasks, together with the data from Exp. 1, collapsed across property types (dangerous and distinct, etc...). [Insert some boringly obvious statistics here.]

### 5.2 Experiment 3b

Experiment 3b differed from Exp. 2 only in that there was only one property type: accidental or temporary properties.

#### 5.2.1 Participants

We recruited 40 participants over Amazon's crowd-sourcing platform Mechanical Turk. 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 4 minutes and participants were compensated $0.40.

#### 5.2.2 Procedure and materials

Our procedure[14] was identical to Exp. 2. On each trial, participants were told: "Listed below are 5 kinds of animals that are found on the island." and asked: "What percentage of each kind of animal do you think has [accidental/temporary property]?"

[14]The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/cbg2010-replication/experiment/prior-5.html`

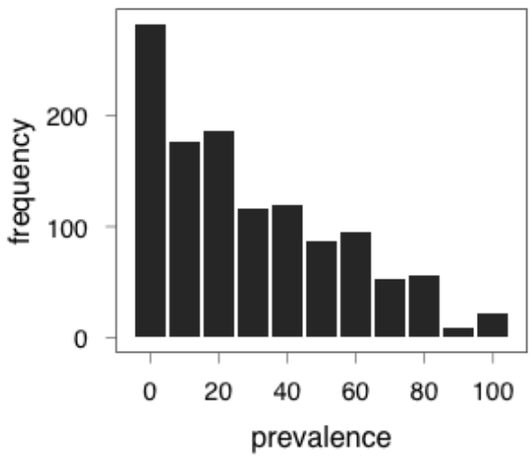



Figure 11: Elicited prior distribution over prevalence of accidental properties.

In Exp. 2, we found high reliability between dependent measures in this prior elicitation paradigm. In this experiment, we ran only the condition with 5 animal-kinds per trial with a slider varying from 0–100 for each animal.

#### 5.2.3 Results

The elicited prior over prevalence for accidental properties was markedly different than the priors over the biological properties. Most importantly, there was no special status given to 100% prevalence. [What else to say about this...]

### 5.3 Modeling results

We use the priors elicited in Exp. 3b to test if our model, like participants, infers a lower overall prevalence when interpreting a generic statement about accidental properties. Using the same bootstrapping method as describe in Section 4.3, we generated model predictions for the truth conditions and implied prevalence tasks using the empirical priors elicited in Exp. 3b. Figure 9 (right) shows the predicted truth conditions and implications for generic statements about accidental properties. Like the human data, the model predictions much weaker implications for generics about accidental properties. This asymmetry goes away and begins to reverse.

Thus, we can see how the shape of the prior over prevalence is critical to fostering the strong implications of generic statements about biological properties. When properties with different shaped priors are under discussion (particularly, priors without large mass at 100%), the asymmetry between truth conditions and implications can disappear or even reverse.

## 6 Experiments 4: Dangerous properties

Our model explains the flexibility of truth conditions in terms of the prior distribution over prevalence. One potential challenge to this is that properties which are not necessarily *distinctive* still garner acceptability as generic statements. For example, CBG found that *dangerous* properties alone increased the proportion of "true" responses to the generic. Here, we sought to replicate these findings, and observe if the prior distribution over *dangerous* properties was in fact similar to the prior distribution over *distinctive* properties. That is, are *dangerous* generics endorsed by virtue of the fact that dangerous properties are distinctive properties?

### 6.1 Experiment 4a: truth conditions

Experiment 4a differed from Exp. 1a only in that the three property types tested were *dangerous* (e.g. "These feathers are as sharp as needles and can easily get lodged in you, causing massive bleeding"), *distinctive* (e.g. "No other

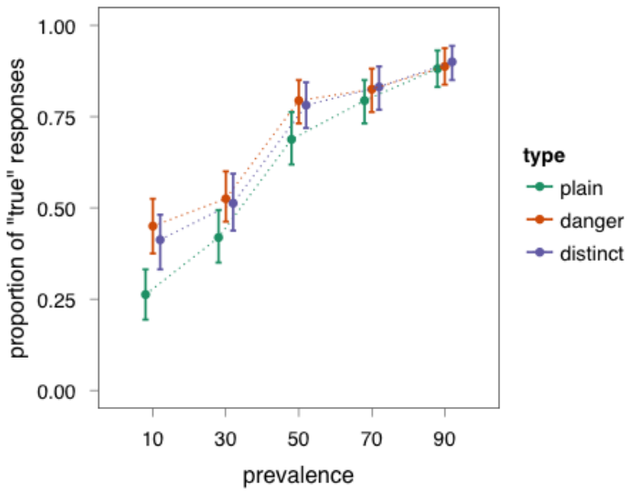



Figure 12: Replication of CBG Exp. 4. Truth conditions for dangerous and distinctive properties (separately) are more relaxed than for plain properties.

animals have these kinds of feathers"), and *plain* (no additional information).

### 6.1.1 Participants

We recruited 80 participants over Amazon's crowd-sourcing platform Mechanical Turk. 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 $0.50.

### 6.1.2 Procedure and materials

All participants were assigned to the *truth conditions* task [15].

We used the same materials as CBG (available in their Appendix). The materials used were 30 novel animal categories (e.g. lorches, morseths, blins) each paired with a unique property. Properties were made by pairing a color with a body-part (e.g. purple feathers, orange tails). Each participant saw 30 unique animal-property pairs: 10 of each of the 3 types (*dangerous*, *distinctive*, *plain*). The 10 items of each property-type were randomly paired with 1 of 5 "prevalence levels": $\{10, 30, 50, 70, 90\}\%$; thus, each prevalence level appeared 2 times per type.

On each trial, participants saw a prevalence statement and type statements (*dangerous*, *distinctive*, *plain*; illustrated above). Participants were then asked "Is the following sentence true or false?", below which was presented the associated generic (e.g. "Lorches have purple feathers") and "True" and "False" radio buttons.

### 6.1.3 Results

Results are shown in Figure 12. We entered participants' truth judgments into a mixed effects logistic regression with random by-item and by-participant effects of intercept and fixed effects of prevalence and type as well as their interaction[16]. Our results replicated the finding of CBG that the generic statements were endorsed more with dangerous properties and with distinctive properties than with plain properties ($\beta_{danger} = 0.85; SE = .18; z = 4.69; p < .001; \beta_{distinct} = 0.83; SE = .18; z = 4.53; p < .001$). There was also an interaction between prevalence level and type such that the generic was endorsed more with *dangerous* properties than with plain properties at lower prevalence levels ($\beta = .02; SE = .007; z = 2.76; p < 0.01$). There was a trending interactive effect for the *distinctive* properties to be endorsed *more* than the plain properties at low prevalence levels ($\beta = .01; SE = .007; z = 1.67; p < .1$).

[15] The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/cbg2010-replication/experiment/experiment-15.html`

[16] This was the maximal mixed-effect structure supported by the data.

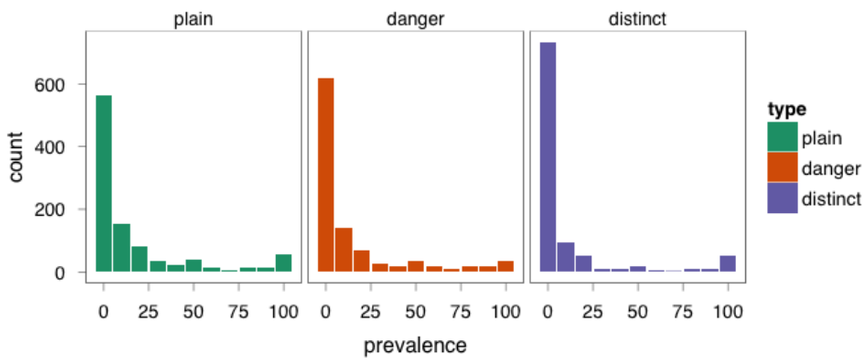



Figure 13: Elicited prior distribution over plain, dangerous, and distinctive properties.

## 6.2 Experiment 4b: prior elicitation

Experiment 4b differed from Expt. 2 only in that the the dangerous and distinctive category was broken up into two categories (*dangerous* and *distinctive*, separately). The only other category tested were "plain" properties (no additional information provided).

### 6.2.1 Participants

We recruited 100 participants over Amazon's crowd-sourcing platform Mechanical Turk. 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 4 minutes and participants were compensated $0.40.

### 6.2.2 Procedure and materials

Our procedure[17] was identical to Exp. 2. On each trial, participants either read information about the property type (*dangerous* or *distinctive*) or no additional information (*plain*).

In addition to the information about property type, participants were told: "Listed below are 10 kinds of animals that are found on the island." and asked: "What percentage of each kind of animal do you think has [property]?" In Exp. 2, we found high reliability between dependent measures in this prior elicitation paradigm. In this experiment, we ran only the condition with 10 animal-kinds per trial with a free response restricted to be a number between 0–100 for each animal.

### 6.2.3 Results

Experiment 4b recovered the shape of biological property prevalence distributions (Figure 13). Hartigans' Dip Test for Unimodality was highly significant for each of the prior distributions ($D = 0.062, 0.039, 0.054$ for types *plain*, *dangerous*, and *distinctive*, respectively; $p < 0.0001$ for each), and thus the distributions are at least bimodal. The means of these three distributions are distinct and ordered (bootstrapped 95% confidence intervals in parentheses): $\mu_{plain} = 16.6\%(14.9, 18.3), \mu_{dangerous} = 14.3\%(12.8, 16.0), \mu_{distinctive} = 11.1\%(12.7, 9.5)$. The medians of these three distributions were all significantly different from one another, evidenced by pair-wise Mann-Whitney U tests (*dangerous* vs. *plain*: $W = 527732.5$; *distinctive* vs. *plain*: $W = 596451.5$; *distinctive* vs *dangerous*: $W = 570639.5$; all $p < 0.05$). Finally, the distributions themselves were all significantly different from one another, by Kolmogorov-Smirnov tests (*dangerous* vs. *plain*: $D = 0.062$; *dangerous* vs. *distinctive*: $D = 0.142$; *distinct* vs *plain*: $D = 0.197$; all $p < 0.05$). In sum, the elicited prior distributions are all at least bimodal, have different central tendencies, and are all distinct in shape.

[17] The experiment in full can be viewed at `http://stanford.edu/~mtessler/experiments/generics/cbg2010-replication/experiment/prior-5.html`

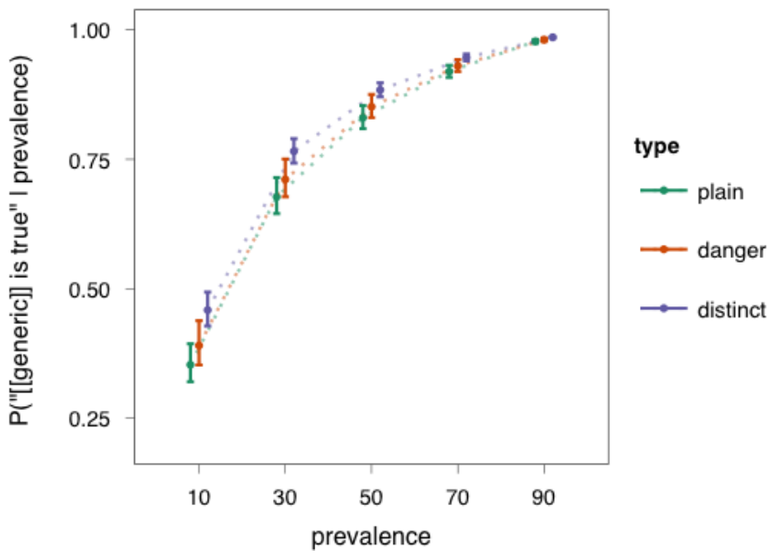



Figure 14: Model predictions for truth conditions of dangerous, distinctive, and plain properties. Error bars denote bootstrapped 95% confidence intervals.

### 6.3 Modeling results

We followed the same model analysis as in Section 4.3. We bootstrapped the model predictions by resampling (with replacement) the empirical prior judgments of Expt. 4b. The model predictions for the truth conditions of separate dangerous and distinct (as well as plain) properties are shown in Figure 14. The model predicts an overall higher proportion of "true" responses to the generic at low prevalence levels for *distinctive* properties relative to *plain* properties. It is unclear whether or not the model also predicts a reliable overall higher proportion of "true" responses at low prevalence levels for *dangerous* properties relative to *plain* properties.

## 7 General discussion

The lifted-threshold RSA model presented in this paper takes a generic statement to be vague. "X has Y" means "many X have Y, *relative to other categories, the vast majority of which, very few or none of the individuals in those categories have Y*". The model predicts overall more nuanced truth conditions for generics of properties that have prevalence distributions which place significant probability mass at or near 0% and significant probability mass at some prevalence greater than 0%. By reasoning pragmatically about an uncertain threshold, the model is able to arrive at gradable and context-sensitive predictions about the truth conditions and implications of generic statements.

We used a Bayesian data analytic model to make a "backward prediction" about the underlying prevalence prior by way of the observed experimental generics data and the lifted-threshold RSA model. We verified this prediction and then incorporated the empirical prior into the model, reconfirming the original fit. We showed how our model predicts a dissipation (or even reversal) of the asymmetry between truth conditions and implications of generics as first explored by Cimpian et al. (2010). Finally, we explored how far prevalence alone could take us, by measuring the prior for just *dangerous* properties. We found that while *dangerous* properties were interpreted as more distinctive and plain properties, this difference only led to marginal increases in the predicted "true" responses in the truth conditions task.

The observation that the predictions for the truth conditions task were *qualitatively* correct but *quantitatively* not as large as the observed behavior data gives rise to at least two possible explanations. First, it is possible that our technique for measuring the prior distribution over prevalence is too crude to account for such small differences in prevalence. It was not guaranteed that differences in the truth conditions tasks between truth judgments at different prevalence levels would actually show up as important differences in the prior elicitation task. Additionally, the measurements

made in the truth conditions task might also be too coarse: In these experiments, truth conditions are measured by categorical judgments (*true* or *false*) at prevalence intervals of 20%. It's possible that more graded judgments at finer prevalence scales would be helpful in mediating this. However, it's also possible that with an uncertain threshold for the generics, participants will adapt to whatever prevalence sampling an experimenter presents. At the end of the day, the truth conditions task is interested in measuring the prevalence level at which the generic is true. It's likely that other, convergent measures would be helpful in assessing this.

A second possibility for the seeming failure to account for the magnitude of the effect in truth judgments for different property types is because prevalence alone is not being used as the only measuring stick for assessing the generic. It's possible that something about the salience of the property (e.g. instantiated in its dangerousness) leads ones to be more flexible in its usage(Leslie, 2008). This explanation could be cashed out in a model of nonliteral language like the one in presented by Kao et al. (2014). The idea here would be that using a generic to describe a salient (e.g. dangerous) property of a kind would be like using hyperbolic language. This type of generic, then, would convey the idea of prevalence in addition to an affective message, like "be careful". We leave for future work exploring this type of explanation in more detail.

## 7.1 The parallel problem of generic identification

In this article, we have focused exclusively on the problem of generic meaning, using sentences with bare plural construction. The bare plural construction is likely the easiest for generic readings. For example, in the work of Prasada, Khemlani, Leslie, and Glucksberg (2013), the overall "true" responses for statements in the bare plural construction (regardless of content) was appreciably higher than all other sorts of constructions associated with generic meaning. In a way, bare plurals are only read generically.

Stepping outside of bare plurals, one runs into a problem parallel to that of generic meaning: generic identification. There is no 'generic operator' that signifies a particular sentence is a generic. Sentences of the same morphosyntactic type can take both generic and non-generic readings. For example, "The bird is in the garage" cannot be read generically while "The bird is a warm-blooded animal" can. Sensitivity to contextual and morphosyntactic cues begins early on (Cimpian & Markman, 2008). It's agreed that world-knowledge must exert influence at some level, e.g. to understand the different readings in "A horse is sick" vs. "A horse is vegetarian" (S. A. Gelman, 2004). It's likely the sort of knowledge used in the lifted-threshold RSA model (i.e. knowledge about the distribution of prevalence) would be useful in the problem of generic identification.

## 7.2 Generics in learning

In this work, we have found that differences in the distributions of prevalence can explain differences in truth conditions for different types of properties. An open question for this line of work is how do children come to have different priors for different property types in the first place? It's estimated that generics account for 4% of all utterances addressed to preschool-age children in everyday contexts (S. A. Gelman, Goetz, Sarnecka, & Flukes, 2008) and that 2-3 year old children comprehend generic statements (Cimpian, Meltzer, & Markman, 2011; S. a. Gelman & Raman, 2003). If interpreting a generic requires knowledge about the distribution of the prevalence of the property, where do children learn that from? One answer might be that learn it at the same time that they learn the meaning of the word in the first place (Frank, Goodman, & Tenenbaum, 2009). Further work should explore the learning problem in terms of a joint inference about the possible category in reference and the still-being-learned meaning of words.

### 7.3 Conclusion

We have explored and demonstrated the viability of a scalar semantics for generics when coupled with a sophisticated pragmatics. A lower-bound threshold on prevalence—the probability of the property given the category—is inferred as part of pragmatic interpretation, yielding vague and context sensitive meanings. We formalized reasoning about the threshold in a lifted-threshold Rational Speech Acts model. This model predicted graded truth judgements and an asymmetry between truth and prevalence judgments. It also naturally accommodates the role of context, explaining these effects as the result of variation in the prevalence prior.

Generics are ubiquitous in natural language. It might seem paradoxical, then, that the semantics of generic statements are underspecified. Why should vague language get so much usage? One possibility is apparent in the lifted-variable RSA model: generic language provides interlocutors with the flexibility to convey rich meanings, which are easily understood in context. Generics are vague, but predictable and useful. 

## References


Brandone, A. C., Cimpian, A., Leslie, S.-J., & Gelman, S. a. (2012). Do lions have manes? For children, generics are about kinds rather than quantities. *Child development*, *83*(2), 423–33. Retrieved from `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3571626\&tool=pmcentrez\&rendertype=abstract` doi: 10.1111/j.1467-8624.2011.01708.x

Brandone, A. C., Gelman, S. A., & Hedglen, J. (2014). Children's Developing Intuitions About the Truth Conditions and Implications of Novel Generics Versus Quantified Statements. *Cognitive science*, 1–28.

Carlson, G. N. (1977). *Reference to kinds in english*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

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.

Cimpian, A., & Cadena, C. (2010). Why are dunkels sticky? Preschoolers infer functionality and intentional creation for artifact properties learned from generic language. *Cognition*, *117*(1), 62–68. Retrieved from `http://dx.doi.org/10.1016/j.cognition.2010.06.011` doi: 10.1016/j.cognition.2010.06.011

Cimpian, A., & Markman, E. M. (2008). Preschool children's use of cues to generic meaning. *Cognition*, *107*, 19–53. doi: 10.1016/j.cognition.2007.07.008

Cimpian, A., Meltzer, T. J., & Markman, E. M. (2011). Preschoolers' use of morphosyntactic cues to identify generic sentences: Indefinite singular noun phrases, tense, and aspect. *Child Development*, *82*(5), 1561–1578. doi: 10.1111/j.1467-8624.2011.01615.x

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Cohen, A. (1999). Generics, Frequency Adverbs, and Probability. *Linguistics and Philosophy*, *22*.

Danovaro, R., Dell'Anno, A., Pusceddu, A., Gambi, C., Heiner, I., & Mobjerg Kristensen, R. (2010). The first metazoa living in permanently anoxic conditions. *BMC Biology*, *30*(8).

Degen, J. (2015). Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 1–48.

Degen, J., & Goodman, N. D. (2014). Lost your marbles? the puzzle of dependent measures in experimental pragmatics. In *Proceedings of the thirty-sixth annual conference of the Cognitive Science Society.*

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084).

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*.

Gelman, S. A. (2004). Learning words for kinds: Generic noun phrases in acquisition. In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon* (p. 445-484). MIT Press.

Gelman, S. A., Goetz, P. J., Sarnecka, B. W., & Flukes, J. (2008). Generic Language in Parent-Child Conversations. *Language Learning and Development*, *4*(1), 1–31. doi: 10.1080/15475440701542625.Generic

Gelman, S. a., & Raman, L. (2003). Preschool children use linguistic form class and pragmatic cues to interpret generics. *Child development*, *74*(1), 308–325.

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., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantic theory, 2nd edition.* Wiley-Blackwell.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition..

Grice, H. P. (1975). Logic and conversation. In *Readings in language and mind.* Blackwell.

Kao, J. T., Wu, J., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*.

Khemlani, S., Leslie, S. J., & Glucksberg, S. (2009). Generics, prevalence, and default inferences. In *Proceedings of the 31st annual conference of the cogntiive science society.* Austin, TX.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a bayesian model of interpretation. *Synthese*.

Lassiter, D., & Goodman, N. D. (to appear). How many kinds of reasoning? inference, probability, and natural language semantics. *Cognition*.

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

Leslie, S.-J. (2008, July). Generics: Cognition and acquisition. *Philosophical Review*, *117*(1).

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. The MIT Press.

Luce, D. R. (1959). *Individual choice behavior: a theoretical analysis*. Wiley.

Prasada, S., Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2013, March). Conceptual distinctions amongst generics. *Cognition*, *126*(3), 405–22. Retrieved from `http://www.ncbi.nlm.nih.gov/pubmed/23291421` doi: 10.1016/j.cognition.2012.11.010

Tessler, M. H., & Goodman, N. D. (2014). Some arguments are probably valid: Syllogistic reasoning as communication. In *Proceedings of the thirty-sixth annual conference of the Cognitive Science Society.*