American Psychological Association
Psychological Review

0033-295X/19/$12.00
2019, Vol. 126, No. 3, 395–436
http://dx.doi.org/10.1037/rev0000142

The Language of Generalization

Michael Henry Tessler and Noah D. Goodman
Stanford University

Language provides simple ways of communicating generalizable knowledge to each other (e.g., "Birds fly," "John hikes," and "Fire makes smoke"). Though found in every language and emerging early in development, the language of generalization is philosophically puzzling and has resisted precise formalization. Here, we propose the first formal account of generalizations conveyed with language that makes quantitative predictions about human understanding. The basic idea is that the language of generalization expresses that an event or a property occurs relatively often, where what counts as relatively often depends upon one's prior expectations. We formalize this simple idea in a probabilistic model of language understanding, which we test in 3 diverse case studies: generalizations about categories (generic language), events (habitual language), and causes (causal language). We find that the model explains the gradience in human endorsements that has perplexed previous attempts to formalize this swath of linguistic expressions. This work opens the door to understanding precisely how abstract knowledge is learned from language.

Keywords: genericity, generalization, generics, pragmatics, semantics

Knowledge that extends beyond the present context is crucial to thrive in our open-ended, dynamic world. Yet, such knowledge can be difficult to extract from the environment: The relevant observations may be costly (e.g., learning that a plant is poisonous) or rare (e.g., understanding that lightning strikes tall objects). Fortunately, we are not limited to acquiring generalizations on our own; language allows us to communicate generalizations to each other. By sharing generalizable knowledge, we flourish collectively without individually needing to taste potentially poisonous plants or personally witness lightning strikes. Being able to flexibly communicate generalizations from one generation to the next supports the faithful transmission of knowledge necessary for culture to cumulatively evolve (Henrich, 2015; Tomasello, 1999).

The *language of generalization* covers a diverse swath of natural language expressions. Generic language conveys generalizations about categories (e.g., "Dogs have four legs"; Carlson, 1977; Cohen, 1999; Leslie, 2007; Nickel, 2008) and is the most well-studied case of generalizations in language.[1] In contrast to statements about concrete individuals (e.g., "Rufus has four legs"), generic statements refer to inherently unobservable categories (e.g., the category of dog) and convey information that extends beyond the present context, a fact that children as young as two appreciate (Cimpian & Markman, 2008). Simple events (e.g., "John ran yesterday") can be generalized into habitual sentences (e.g., "John runs"), and even events of complex inferential types such as actual causal events (e.g., "The fire caused the smoke") can be described in generalization (e.g., "Fire causes smoke").

Understanding the language of generalization is a project with far-reaching implications. The language of generalization is ubiquitous in everyday conversation, is found in every language (Behrens, 2005; Carlson & Pelletier, 1995), and conveys rich meanings, impacting motivation (Cimpian, Arce, Dweck, & Markman, 2007), transmitting stereotyped beliefs about social categories (Rhodes, Leslie, & Tworek, 2012), and making meaning from experience (Orvell, Kross, & Gelman, 2017). It is highly prevalent in child-directed speech (Gelman, Goetz, Sarnecka, & Flukes, 2008) and its

This article was published Online First February 14, 2019.

Michael Henry Tessler and Noah D. Goodman, Department of Psychology, Stanford University.

This article is based on a Ph.D. dissertation submitted to Stanford University. The research was supported by National Science Foundation Graduate Research Fellowship DGE-114747 awarded to Michael Henry Tessler and DARPA PPAML grant FA8750-14-2-0009 and an Alfred P. Sloan Research Fellowship to Noah D. Goodman. A portion of this article (Case Study 2) appeared in the proceedings of the 38th Annual Meeting of the Cognitive Science Society, Philadelphia, 2016. We are deeply grateful to Ellen Markman, Gregory Scontras, Thomas Icard, Michael C. Frank, and Sophie Bridgers for extensive conversations and support for this work. In addition, the work was greatly improved through countless discussions with members of the Computation and Cognition Lab and the Language and Cognition Lab at Stanford, as well as the participants of *A Generic Workshop* at the Center for the Study of Language and Information (CSLI) at Stanford in May 2017. Finally, we thank Michael Franke and Joshua Hartshorne for helping us clarify our main arguments and improving the quality of the manuscript. All models, analyses, data, and links to experiments used in this article can be found at https://github.com/mhtess/genlang-paper.

Correspondence concerning this article should be addressed to Michael Henry Tessler, who is now at the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Mass Avenue, Building 46-3027 Cambridge, MA 02139. E-mail: tessler@mit.edu

[1] Some writers refer to the all of the *language of generalization* as "generic language" or generics. In both the empirical and theoretical literatures, however, analysis and experiments often focus only on generalizations about categories. We use the more narrow-scoped terms (e.g., generics, habituals, and causals) to highlight the diversity of semantic types being predicated.

ability to refer to abstractions beyond the present context suggests its centrality to the growth of conceptual knowledge (Gelman, 2004).

Despite its ability to convey abstract knowledge, its ubiquity in discourse, and its relative morphosyntactic simplicity, the language of generalization displays subtle context-sensitivities that make it difficult to formalize. "Robins lay eggs" sounds true and "Robins are female" does not. Yet, in each case, only 50% of the category has the property (i.e., only the females lay eggs). "Mosquitos carry malaria" sounds true despite malaria being present in only a tiny fraction of mosquitos. Even more perplexing: "Supreme Court Justices have even social security numbers" is thought to be intuitively a false generalization even if it were the case that on the current bench, nine out of nine justices had even social security numbers (i.e., when the sentence "All Supreme Court Justices have even social security numbers" is true; Cohen, 1999). Similar context-sensitivity can be observed with habituals: "Mary climbs mountains" could imply a few mountain climbs a year for Mary, but "John runs" would be infelicitious if John went for a run three times last year.

These observations have led some to conclude that the literal meaning of a generic statement (and by analogy, other generalizations in language) involves more than just the number of members of the category who have the property, otherwise known as the *prevalence* of the feature in the category. Theorists have thus argued that generics are not treatable by the standard tools of truth-functional semantics (Montague, 1973), but rather should be thought of semantically as a direct, linguistic manifestation of abstract relations between kinds and properties (Leslie, 2008; Prasada, 2000; Prasada, Hennefield, & Otap, 2012). For example, the statements "Bishops move diagonally" or "The Speaker of the House succeeds the Vice President" are true not because of a tendency on behalf of instances of a category to actually uphold the property, but rather the existence of a conceptual relationship (e.g., what it means, in the game of chess, to be a bishop). True generics can be supported by different underlying types of category—property relations (e.g., principled vs. statistical connections; Prasada & Dillingham, 2006) and, thus, support qualitatively different inferences (e.g., "Being striped is one aspect of being a tiger" is generally endorsed, while "Carrying malaria is one aspect of being a mosquito" is not, even while "Tigers are striped" and "Mosquitos carry malaria" are both intuitively true; Prasada, Khemlani, Leslie, & Glucksberg, 2013). This conceptual view of generics has been influential in psychology, because it predicts qualitative differences between different kinds of generics.

Insofar as there is a single class of linguistic expressions that convey generalization, however, there should be something common to them all: a literal meaning that unifies generic, habitual, and causal language. In this article, we propose such a semantic core based on prevalence and formalized using the tools of truth-functional semantics (Cohen, 1999; Montague, 1973). A semantics based on prevalence will not be enough to capture subtle sensitivities to context that the language of generalization exhibits. We propose that the meaning of generalizations is underspecified or *vague*, and that listeners derive a more precise interpretation in context using probabilistic world knowledge.

The fact that generics are vague does not preclude them from being treated with formal models. We draw upon the tools of Bayesian models of cognition (Tenenbaum, Kemp, Griffiths, & Goodman, 2011) to formalize the vagueness and context-dependence of generic language (Frank & Goodman, 2012; Goodman & Lassiter, 2015). The Bayesian model separates the semantics of an utterance conveying a generalization from the world knowledge a listener would use to interpret the utterance, a key theoretical advancement beyond previous accounts. This formal model is the first of its kind to make quantitative predictions about human understanding of the language of generalization.

The article is organized as follows. In the next section, we describe our computational framework for interpreting generalizations and our precise model for endorsing such statements (i.e., a model of truth judgments). To illustrate how the model works, we then work through a number of standard examples from the linguistics literature that any model of generics should be able to accommodate. The third section discusses the relationship of our model to previous accounts of generics from the linguistics and philosophical literatures. These theoretical sections are followed by three empirical case studies: generalizations about categories (*generic language*), events (*habitual language*), and causes (*causal language*). In the case study of generics, we measure relevant background knowledge and prevalence to predict endorsements of familiar generic statements (e.g., "Robins lay eggs"). In the study of habituals, we measure background knowledge but manipulate the prevalence or frequency of the event to predict endorsements of habitual statements about novel agents (e.g., "John runs" given that he has run a certain number of times in the past). In this case study, we also further examine the nature of the relevant probabilities for endorsing generalizations, asking whether it matters how often John has run in the past (past frequency) or how often a speaker expects him to run in the future (predictive probability). Finally, in our last case study, we manipulate background knowledge to show its causal influence on endorsing generalizations about novel causal events (e.g., "Herb X makes wugs sleepy"). We compare our model to two previously articulated quantitative models of generics as well as a lesioned-version of our model. In all cases, we find a very strong agreement of our model's predictions to human elicited endorsements, where the simpler models fall short. We conclude our article by clarifying some of the theoretical claims of the model and discuss open-questions for this approach.

Computational Framework

Generalizations are used to make predictions about events or properties of instances that an agent has yet to experience (Hume, 1738). People readily predict that the next dog they encounter will have four legs, drinking another cup of coffee will cause jitters, and a new day will find the people that we know doing what they habitually do. In each case, we assign a specific exemplar x to a category k, and make a prediction that it will have feature f. This prediction can be described by a conditional probability: $P(x \in f \mid x \in k)$, the probability that x will have f given that it is in k (formally, an instance of k will be in the set of things that have f), which we will refer to as the *prevalence* and for convenience write as $p = P(f \mid k)$. The prevalence p is a *prevalence in the mind*, a latent belief that a future instance of a category would have a particular property; others might call this projectibility (Goodman, 1955). The targets of our predictions can vary widely: They may be objects (a dog), events (an instance of coffee-drinking), or more ad hoc types (a person on a particular day). The properties also may vary (e.g., having fur, causing jitters, and going to the gym). Yet, the mathematical description of the inductive belief is always given by a probability $P(f \mid k)$.

Probability is a useful representation for human generalization from observations (Shepard, 1987). If you observe several *wugs* (a novel category) that have two legs, you might infer that all wugs have two legs. However, not all properties have such strong projectibility: Seeing a wug with broken wings tells you comparatively less about other wugs having broken wings (Nisbett, Krantz, Jepson, & Kunda, 1983). Abstract, potentially domain-specific beliefs about the projectibility of different properties can be represented by a probability distribution over the prevalence $P(p)$ (Kemp, Perfors, & Tenenbaum, 2007). By assuming some generative process that could produce one's observations o—a likelihood function $P(o \mid p)$—Bayes's theorem provides the mathematically correct way to update one's prior beliefs from observations: $P(p \mid o) \propto P(o \mid p) \cdot P(p)$ (Tenenbaum & Griffiths, 2001).

Observational data are not always available, however. Instead, we must listen to others to learn about properties that are costly to observe (e.g., *staring at the sun makes you go blind*), events that are statistically unlikely (e.g., *lightning strikes tall objects*), or any aspect of the world that we have yet to experience. Fortunately, language provides simple ways of communicating generalizations.

Communicating Generalizations

The language of generalization is easy to spot when a property, which could apply to an individual, is predicated of a category (e.g., "Dogs have four legs": *has four legs* could apply to an individual as in "Rufus has four legs"). In the semantics literature on generics, bare plural sentences of this kind are sometimes described as *characterizing sentences* in contrast with *kind-denoting* sentences where the property can only meaningfully apply to the category as a whole (e.g., "Dinosaurs are extinct"; *extinct* cannot apply to an individual dinosaur). Generalization can also manifest when describing instances of an individual (*habitual language*; e.g., "John smokes"; Carlson, 1977, 2005); in this case, the particular instance being generalized is an instance of an individual (e.g., John at a particular moment in time), which also permits predication (e.g., "John smoked yesterday after dinner"). Verbs like *causes* or *makes* also seem to convey generalization, in this case, about an instance of an actual causal event (*causal language*; e.g., "Fire causes smoke"). Psychologists, linguists, and philosophers have long studied the language of generalization, as it appears very simple (e.g., syntactically) yet its meaning is difficult to formalize.[2]

The basic intuition behind our account is that before a listener hears a novel generalization such as "Alligators grow to be 10-feet long," they do not know how widely distributed the property to be in the category, including whether or not it is present at all. The utterance provides a vague sense of how strongly the generalization applies (e.g., how many alligators grow to be 10-feet long), which the listener derives from their knowledge of how the property (*growing to be 10-feet long*) is distributed among other categories (e.g., other animals). The decision of whether or not to endorse the generalization is that of a speaker reasoning about how well the utterance would align their interlocutor's beliefs about the prevalence of the feature in the category with those of their own. We formalize this intuition in a truth-conditional semantics incorporated into a Bayesian model of belief updating.

Interpretation model. Our model of interpreting the language of generalizations has three conceptual components: Probability, vagueness, and context. If generalization from observations can be described by a probability p, it is natural to posit that same construct will be at the heart of a semantic theory of the language of generalization (Ingredient 1: Probability). In semantics, belief updating generally passes through Boolean *truth values* (Montague, 1973). The simplest way to derive a Boolean from a scalar quantity like probability is via a *threshold semantics*: The utterance is true if the relevant scalar value is above a threshold. For example, the literal meaning of the sentence "Some dogs have four legs" is that there is a nonzero chance that a given dog will have four legs: $[\![some]\!](p) := p > 0$. "Most dogs have four legs" can also be described as a threshold on prevalence (e.g., the chance that a dog will have four legs is greater than 50%): $[\![most]\!](p) := p > 0.5$.[3] Thus, the simplest semantics for a generalization would also be a threshold on the prevalence: $[\![gen]\!](p, \theta) := p > \theta$.

The extreme flexibility of generalizations (e.g., "Mosquitos carry malaria"; "Birds lay eggs" vs. "Birds are female") suggests that no fixed value of the threshold would suffice. Rather than throw out the threshold-semantics, we posit that the threshold is *underspecified* in the literal meaning and is contextually determined in a way analogous to how gradable adjectives like *tall* have contextually determined thresholds (e.g., what counts as tall for a person is different than what counts as tall for a building; Kennedy, 2007; Lassiter & Goodman, 2013). We formalize this underspecification of meaning (Ingredient 2: Vagueness) by putting a probability distribution over θ: $P(\theta)$ (Lassiter & Goodman, 2013, 2015; cf., Qing & Franke, 2014).

Finally, the meanings of linguistic expressions manifest in their capacity to convey information from speaker to hearer (Clark, 1996; Grice, 1975; Levinson, 1995). Thus, the final ingredient to our model is *context* (Ingredient 3), which must be minimally formalized as a listener's prior knowledge about the property $P(p)$ and that we describe in the Worked Examples section below.

Putting these three ingredients (probability, vagueness, and context) together, we get the following model of interpreting generalizations in language:

$$L(p, \theta \mid u) \propto \delta_{[\![u]\!](p,\theta)} \cdot P(\theta) \cdot P(p) \tag{1}$$

We denote this probabilistic interpretation model by L to indicate that it is modeling a listener updating their beliefs according to the truth-functional meaning of an utterance u. Formally, the truth-functional meaning is represented by the Kronecker δ function $\delta_{[\![u]\!](p,\theta)}$ that returns 1 when the utterance is true (i.e., when $p > \theta$) and 0 otherwise.

[2] We further distinguish the problem of formalizing a *meaning* for the language of generalization from the problem of *identifying* generalizations. Syntax alone is neither necessary nor sufficient for a listener to know that the sentence conveys a generalization (e.g., indefinite singulars can encode generalizations: "A dog has four legs" and bare plurals may not: "Dogs are on my front lawn"). Our analysis, thus, begins once a sentence has been disambiguated as conveying a generalization.

[3] These definitions concern the standard semantic truth conditions of quantifiers, not their pragmatic interpretations (e.g., that "some" often implies "not all").

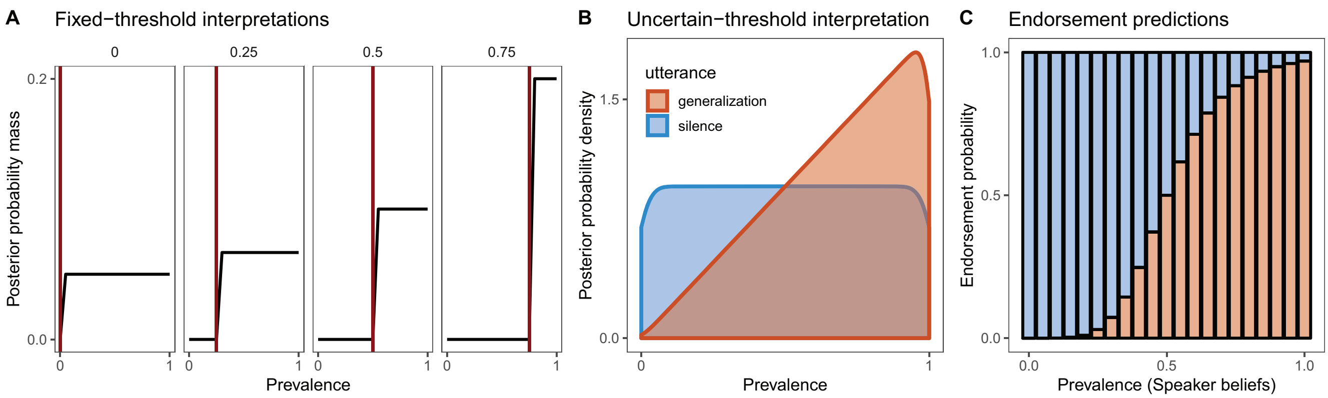

Figure 1. Computational model behavior assuming a uniform prior over prevalence. (A) Interpretation model posteriors assuming different fixed thresholds (facets, also vertical lines). High thresholds rule out more world-states (prevalence levels), leaving fewer world-states among which to distribute the full probability mass. (B) Generic interpretation model averages over all thresholds to return a posterior distribution that favors higher prevalence levels in a graded manner. (C) Endorsement model predicts higher rates of endorsements as prevalence levels increase. See the online article for the color version of this figure.

$$\delta_{[\![u_{gen}]\!](p,\theta)} \propto \begin{cases} 1 & \text{if } p > \theta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Following Lassiter and Goodman (2015), we formalize the vagueness in the meaning of generalizations $P(\theta)$ as a uniform distribution over the support of $P(p)$.

With these ingredients, the literal interpretation model (Eq. 1) computes a posterior interpretation distribution on prevalence by considering different possible thresholds θ. Figure 1 shows the model behavior assuming a uniform prevalence prior $P(p)$, representing a purely abstract property for which an interpreter has no substantive prior knowledge. If the threshold were to be very high (e.g., $\theta = 0.75$), only the highest prevalence levels would be consistent with the utterance. As the threshold decreases, more and more prevalence levels become compatible with the threshold (Figure 1A). The interpretation model weights the different thresholds by the probability that each is true, resulting in a posterior interpretation distribution that favors higher prevalence levels (Figure 1B). This interpretation distribution captures the intuition that generalizations imply high prevalence, though with substantial uncertainty about the precise quantity.

The model in Eq. 1 describes a general mechanism of belief-updating via an uncertain threshold semantics. The semantic scale is determined by the prevalence prior $P(p)$. For generalizations about categories (i.e., generic statements), $P(p)$ ranges over [0, 1]; p denotes the chance that an instance of a category has a Boolean property. To apply this same model to generalizations about events (i.e., habituals like "John runs"), we use a different, but related, scale: the frequency within a given time window (e.g., number of times John went for a run in the past week).[4] For events then, p is a rate and $P(p)$ ranges over $[0, \infty)$.

Endorsement model. In this article, we are interested in explaining truth judgments (e.g., "Robins lay eggs" is intuitively true; "Robins are female" is not). We model the endorsement task (e.g., *true* vs. *false*; *agree* vs. *disagree*) as a *speaker decision* about whether or not to produce the generalization to a naïve listener; thus, the endorsement model has two alternative utterances: produce the generalization versus a null alternative (Degen & Goodman, 2014; Franke, 2014). For simplicity, we take the null alternative to be an informationless "silent" utterance that is always true: $[\![null]\!](p,\theta)=\text{True}$.[5] This kind of speaker model defined in terms of a listener model formalizes the basic aspects of communicative reasoning and is the simplest instantiation of a Rational Speech Act model (Frank & Goodman, 2012; Goodman & Frank, 2016).

The endorsement model (called S for speaker) decides whether or not to produce the generalization by reasoning about how the listener model (L; Eq. 1) would interpret it:

$$S(u \mid p) \propto \left(\int_{\theta} L(p, \theta \mid u) \mathrm{d}\theta\right)^{\lambda} \tag{3}$$

The goal of the endorsement model is to align the listener's *a priori* beliefs about prevalence (given by the prevalence prior) with the speaker's beliefs about the prevalence of the feature in the referent category p, the *referent prevalence* (e.g., that 50% of robins lay eggs).[6] The endorsement model makes an approximately rational (with degree of rationality λ), information-theoretic decision based on its beliefs about which utterance would best achieve the goal of conveying the referent prevalence p. The semantic threshold θ is necessary for establishing the truth condi-

[4] This measure is closely related to the instantaneous probability density of the event, as the time window gets infinitesimally small.

[5] The null alternative can be realized in at least two other ways: the negation of the generalization (e.g., "It is not the case that Robins lay eggs") or the negative generalization (i.e., "Robins do not lay eggs"). All results reported are similar for these two alternatives, and we use the alternative of the "silent" utterance for simplicity.

[6] A more general version of this model can relax the assumption that the endorser/speaker has access to a specific prevalence p that it wants to communicate. Rather, the endorser may have probabilistic beliefs about the prevalence of the property for the referent category k, which would be represented by a distribution over p: $P(p \mid k)$. In this situation, we would define the endorsement model decision to be with respect to the expected value of the informativity, which integrates over the endorsement model's belief distribution: $S(u \mid k) \propto \exp(\lambda \cdot \mathbb{E}_{p \sim P_k} \ln \int_{\theta} L(p,\ \theta \mid u) \mathrm{d}\theta)$. For the empirical case studies described below, these two versions of the model make almost identical predictions.

tions and deriving an interpretation (used in Eq. 1) but is otherwise a nuisance parameter, which the endorsement model integrates out (note that there is no θ in the left-hand side of Eq. 3). Given that the speaker only has two options (i.e., produce the generalization utterance or stay silent), the endorsement decision comes down to whether or not the referent prevalence is more likely under the prevalence prior distribution $P(p)$ (i.e., the listener's posterior upon hearing silence) or the listener generic interpretation distribution $L(p \mid u)$ given by Eq. 1 (Figure 1C). In the next section, we work through different examples from the linguistics literature, describing the prevalence prior and other relevant model components for each.

Worked Examples

We explore the predictions of the endorsement model in the context of a few worked examples. These examples are taken from the linguistics literature on generics and are statements that have been historically challenging for prevalence-based approaches (see Table 1). We implemented these and all subsequent Bayesian models in the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014). All models, analyses, data, and links to experiments used in this article can be found at https://github.com/mhtess/genlang-paper.

Endorsement predictions depend upon the uncertain threshold model's interpretations, which are highly sensitive to the interpretation model's background knowledge. Background knowledge may be richly structured, reflecting intuitive theories about the underlying causes of different kinds of properties (cf., Leslie, 2007). Our model posits, however, that the only impact of structured knowledge on truth judgments is in their implications for beliefs about prevalence, formalized in terms of a prevalence prior $P(p)$. The prevalence prior reflects a listener's beliefs about the prevalence of the feature in an unknown or unfamiliar category and be conceptualized in a theory-neutral way as a distribution over the prevalence among alternative categories (see Figure 2).[7] For example, the *lay eggs* prevalence prior should be bimodal, with substantial probability mass near zero-prevalence (because many animal categories do not have egg-layers) and a secondary component peaked around 50% (because among the animals with egg-layers, only the female members of the category have the property). Conversely, the distribution over the prevalence of *being female* is unimodal and centered at 50%-prevalence, because almost all animals have female members in that proportion. Figure 2 (row 1) shows these and other intuitive prevalence priors for different properties (we introduce methods for empirically eliciting these priors in the experimental section). We now turn our attention to the endorsement model's predictions.

1. Dogs bark. Often the first observation with generics is that they appear to behave like universal quantifiers (e.g., "All") that permit exceptions: Not all dogs bark, but still "Dogs bark" is true (Carlson & Pelletier, 1995). The uncertain threshold account immediately accommodates this example because there is no single, fixed threshold beyond which a generic statement becomes true. Rather, listeners have uncertainty about the threshold that leads to graded interpretations (e.g., "Dogs bark" means almost all dogs bark; Figure 2, column 1, row 2). The endorsement model then predicts "Dogs bark" is a rather good generic sentence (Figure 2, bottom row) even though there are exceptions to the universal generalization (referent prevalence of barking among dogs roughly 95%, shown with arrow in Figure 2, row 2).

2. Kangaroos have spots. A second observation regarding generics is that their truth conditions cannot be so lenient so as to always convey existential quantification (e.g., "Some"). Indeed, it is intuitively plausible that some kangaroos do have spots but a rational language user might feel awkward to assert the generic "Kangaroos have spots." The model exhibits this same restraint, predicting a very low endorsement probability for this generic sentence (Figure 2, column 2). The reason is that, given the interpretation model's background knowledge about the property, the statement "Kangaroos have spots" will be interpreted like "Dogs bark": it would mean almost all kangaroos have spots. The endorsement model, which believes that very few kangaroos have spots, then would rather not endorse the statement because it would lead to a too-strong interpretation and mislead the listener.

3. and 4. Robins lay eggs vs. Robins are female. One of the most difficult examples for a theory of generics based on prevalence to handle is the intuitive difference in truth value between the statements "Robins lay eggs" and "Robins are female." The former is intuitively true, even though only female robins could lay eggs (and hence, the prevalence is roughly 50%). However, the same implicit restriction to only females does not seem to occur for the latter statement, as "Robins are female" strikes most as strange or false. Why is "Robins lay eggs" a reasonable utterance while "Robins are female" is not?

The prior distributions over the prevalence of both features are shown in Figure 2 (columns 5 and 6). As described above, the priors are different: Many animals have zero egg-layers (0% prevalence), while the vast majority of animal categories have female members in exactly the same proportion (50%). Given this background knowledge, the generic interpretation model returns roughly the same interpretation distribution for each hypothetical utterance: In each case, the model believes roughly 50% have the property. However, only in the case of "Robins lay eggs" does the endorsement model actually assert the generic; it does so because the listener interpretation would be more aligned with the referent prevalence in comparison with the interpretation of the null utterance, which is the prevalence prior. In contrast, the hypothetical generic interpretation of "Robins are female" is not different from the prevalence prior and hence the generic conveys no new information; here, the generic would not be misleading but uninformative, and hence the model predicts an endorsement probability of 0.5. Indeed, previous studies on generic endorsements have found "Robins are female" to not be rated as completely false but rather receive an intermediate endorsement level (i.e., neither true nor false; Prasada et al., 2013).

5. Mosquitos carry malaria. The statement "Mosquitos carry malaria" is intuitively true despite the fact that the vast majority of mosquitos are actually malaria-free. The prevalence prior for *carries malaria* is highly skewed toward low prevalence levels: many animals do not have malaria-carriers among them and even for those that do the prevalence is expected to be quite low. Then, carrying malaria is significantly more true of mosquitos than other

[7] Alternatively, the prevalence prior can be conceptualized as a marginal distribution on prevalence derived from an intuitive theory formalized in a probabilistic language of thought (e.g., Goodman et al., 2015).

Table 1
Example Sentences That a Theory of the Language of Generalizations Should Correctly Predict

Example	Intuitive referent prevalence	Intuitive truth value	Issue
1. Dogs bark	95%	True	Not all dogs bark
2. Kangaroos have spots	5%	False	Some kangaroos could have spots
3. Robins lay eggs	50%	True	Only female robins (50%) lay eggs
4. Robins are female	50%	False	Same number that lay eggs (50%)
5. Mosquitos carry malaria	5%	True	Very few mosquitos actually carry malaria
6. Sharks don't eat people	95%	False	Most sharks don't eat people
7. Mary handles the mail from Antarctica	0 times	True	Consider there has never been any mail from Antarctica
8. Supreme Court Justices have even social security numbers	100%	False	Consider 100% have even social security numbers
9. Elephants live in Africa and Asia	50%/50%	True	Impossible for most to live in Africa and most to live in Asia, and no individual elephant lives in both Africa and Asia

Note. See details of each example in corresponding section.

animals, an intuition that is often arrived at with this example. This kind of behavior is related to the construct of *cue validity*—the probability of the feature given the category for example, $P(\textit{is a mosquito} \mid \text{carries malaria})$—which we return to in the next section. The uncertain-threshold model displays a critical behavior: It can endorse generics when the referent prevalence is very low.

6. Sharks don't eat people. When the prevalence of the feature is very high, it is not necessary for the model to endorse the generic. "Sharks don't eat people" is predicted to be a somewhat strange utterance, despite the fact that the vast majority of sharks do not eat people.[8] Because for almost all animal categories, the prevalence of *not eating people* is almost 100% (very few things eat people), interpreting the statement "Sharks don't eat people" would lead one to believe that no sharks eat people (i.e., 100% do not eat people), which is too strong. It is interesting to note that because of the high referent prevalence, the model is less certain of its decision, predicting an endorsement probability around 0.4 (i.e., somewhat false). We will test this quantitative prediction in Experiment 1 on generic language.

7. Mary handles the mail from antarctica (yet has never had the opportunity). Imagine there is a job in the local bureaucrat's office to handle the mail from Antarctica and this job is assigned to Mary; to date, however, nobody has ever sent mail to the office from Antarctica (Cohen, 1999). In other words, the statement "Mary has handled mail from Antarctica" is false. "Mary handles the mail from Antarctica," however, is still thought to be intuitively true despite zero actual instances of the event. This highlights an important ambiguity in the theoretical commitments of the uncertain threshold model: The endorsement model aims to communicate some referent prevalence p, but does the prevalence represent the actual, objective frequency in the world (e.g., the number of times in the past that Mary has handled mail from Antarctica) or a subjective, predictive degree of belief in the head (e.g., our prediction that were the appropriate situation to arise, Mary would be handling the mail from Antarctica)?

We posit that Mary the Antarctic mail handler is an extreme case of generalizations expressing *predictive* degrees of belief. We expect this context sets up the expectation that any future mail coming from Antarctica will be handled by Mary. Thus, the predictive probability that Mary will handle Antarctic mail, should there be some, is high. Predictive probabilities often track past frequency or actual prevalence in the world, but people's internal models of how the world works can lead the two to diverge. In this case our understanding of Mary's job leads to a strong predictive probability in the absence of past frequency evidence. We explore this question experimentally in Experiment 2 on habituals.

8. Supreme Court Justices have even social security numbers. Imagine that all current Supreme Court Justice has a social security number that was an even number. "Supreme Court Justices have even social security numbers" is still considered false, even though the property holds for exactly 100% of the category (Cohen, 1999). We predict the rejection of even social security numbered Justices is the result of people's intuitive theories guiding their subjective predictive probabilities, which feed into the endorsement model. That is, we predict observers strongly believe there is no causal relation between the evenness of one's social security number and selection for the Supreme Court and, thus, would assign a roughly 50% subjective probability to the next justice having an even social security number. Then, "Supreme Court Justices have even social security numbers" would be rated by our model as similar to "Birds are female," because all professions have roughly the same probability of having employees with even social security numbers. This example is, thus, the conceptual opposite of Mary the Antarctic mail handler, where our internal model of the world led us to a strong degree of belief in the property holding in the future; with the Supreme Court Justices, our internal models lead to a relatively weak degree of belief in the property holding in the future.

9. Elephants live in Africa and Asia. Understanding how a semantic representation *composes* is another important test for a theory of the language of generalization. Nickel (2008) suggests that "Elephants live in Africa and Asia" is troubling for prevalence-based accounts (in particular, majority-quantificational accounts where the generic means *more than half*; Cohen, 1999) because the statement should be semantically equivalent to "Elephants live in Africa and elephants live in Asia." It cannot be the

[8] We create this example as the converse of the classic example "Sharks eat people," predicted to be *true* despite low prevalence (Leslie, 2007). We reverse this example to show how the model deals with a high-prevalence feature for which the generic is predicted to be false or infelicitous.

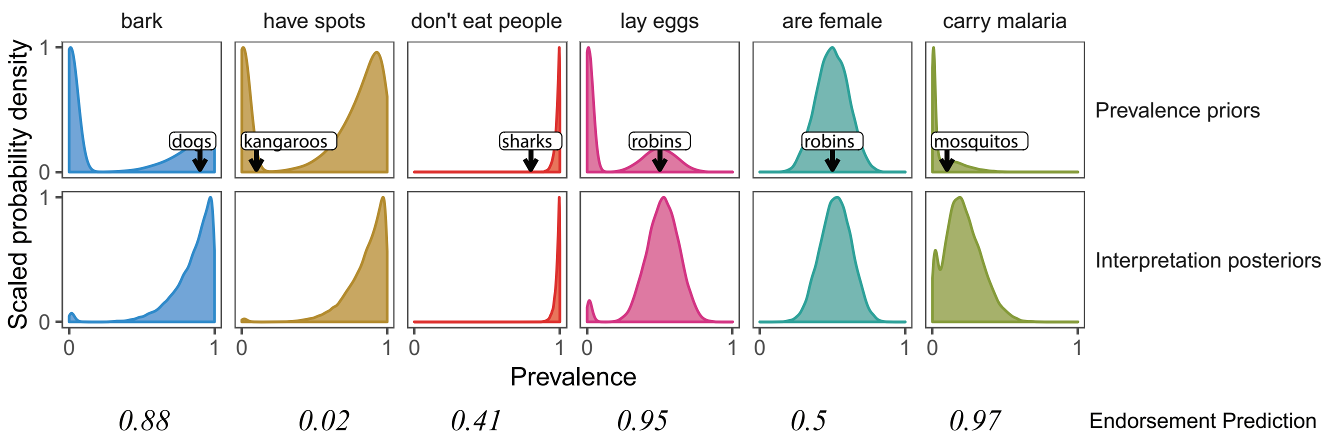

Figure 2. Model simulations assuming different prevalence priors. Top: Prevalence priors for six example features. Shapes of the priors were chosen to intuitively correspond to the properties labeling the distributions. Arrows show referent-prevalence for a target category. Bottom: Interpretation model posterior distributions over prevalence upon hearing a generalization about a novel category. Numbers at bottom correspond to endorsement model predictions for endorsing the generalization for the referent-category whose prevalence is shown in the top facets (e.g., "Mosquitos carry malaria"). See the online article for the color version of this figure.

case that most (more than half of) elephants live in Africa and most (more than half of) elephants live in Asia, unless we are to posit the existence of international elephants (i.e., individual elephants who live part-time in Africa and part-time in Asia), which are intuitively implausible.

Our theory does not provide a fixed semantics for generics, but an uncertain one which can be updated as more information comes in. In fact, with prior knowledge suggesting against the existence of international elephants (Figure 3A), our model interprets "Elephants live in Africa and Asia" as meaning that some elephants live in Africa and that different ones live in Asia (Figure 3C). Of theoretical interest, an incremental parsing of the sentence (i.e., upon hearing only that "Elephants live in Africa") leads our model to believe that most, possibly all, elephants live in Africa (Figure 3B). When the sentence is completed (". . . and Asia"), the model nonmonotonically updates its beliefs to something weaker: some elephants live in Africa and others in Asia (Figure 3C). The flexibility of our model accommodates new evidence that might

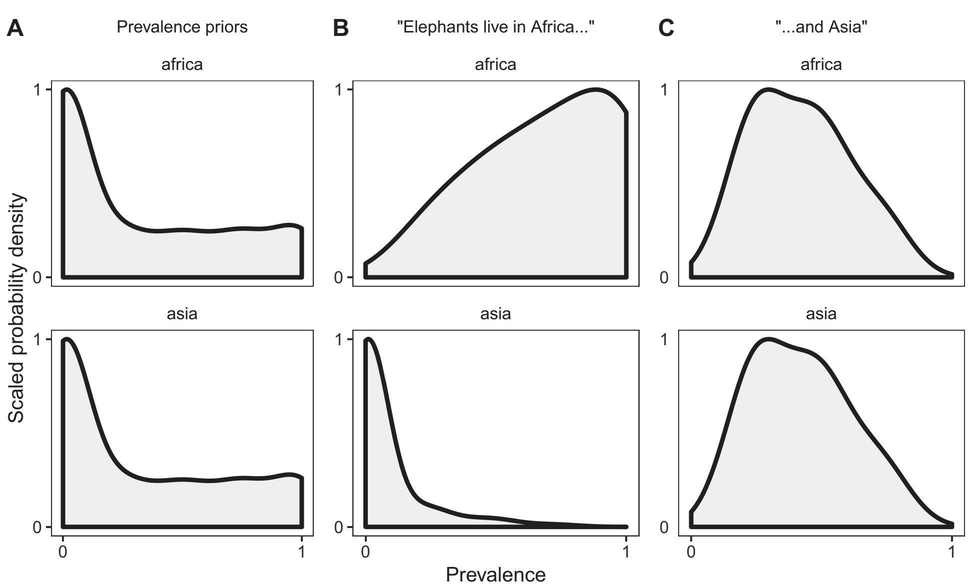

Figure 3. Model interpretation of a conjunctive generic ("Elephants live in Africa and Asia"). (A) Intuitive priors for marginal distributions of prevalence for living in Africa and living in Africa. (B) Partial interpretation upon hearing utterance ("Elephants live in Africa . . ."). (C) Full interpretation upon hearing the end of the sentence ("Elephants live in Africa and Asia"). Uncertainty about the threshold for generic sentences allows the listener to nonmonotonoically update its beliefs about the prevalence of the feature in categories.

otherwise contradict previous utterances because it maintains uncertainty about the precise meaning of the utterance. From a compositional perspective, this is highly desirable behavior but we leave the testing of quantitative predictions of this kind for future work.

Relationship to Other Semantic Theories

Our formal theory builds on and relates to a number of extant theories of generics in both the formal semantics and psychological literatures. The dominant approaches from formal semantics try to describe unified, objective criteria by which to establish if a generic and other generalizations are true or false (i.e., generic *truth conditions*). These views are typically *statistical* in nature in that they appeal to quantification or the statistics of the world (e.g., how many of the kind have the property) to define those unified, objective criteria. Many of these theories rely upon mechanisms outside of the truth conditions to make sense of the extreme flexibility exhibited by generics; unfortunately, these extrasemantic mechanisms are not described in sufficient detail to generate precise, quantitative predictions. Other theorists, primarily in philosophy and psychology, have taken the extreme flexibility of generics to argue against a quantitative theory based on statistics, instead suggesting that abstract, mental representations are directly tied to semantics of generics (e.g., *there is something about being a K which causes it to F*). Statistical and conceptual theories express the major contrasting views of generic language (Carlson, 1995).[9]

Conceptual Accounts

Conceptual accounts try to identify the core meaning of a generalization directly with aspects of conceptual structure. The most influential account in psychology comes from Leslie (2008), which uses the same starting point as our analysis: Generics express generalization. Leslie (2008)'s analysis draws upon insights from the psychological literature on infant generalization to argue that generics tap into an innate, default mechanism that signals the child (or adult) to generalize the property to the kind.

The *default generalization* mechanism has three components: the ability to (a) identify whether or not a feature is a *characteristic* property of a kind (e.g., one characteristic property for an animal would pertain to the mode of locomotion for the animal), (b) identify whether or not a feature is *striking* (e.g., can kill you; "Mosquitos carry the West Nile Virus"), and (c) segment *counterinstances* (instances of the kind that do not have the feature) into positive and negative counterinstances (i.e., instances of the kind that have some relevant alternative feature and instances of the kind that simply lack the feature, respectively).[10] On this account, generics are true if some instances of the kind have the feature when all counterinstances are negative counterinstances and the property is either characteristic or striking. If the property is neither characteristic nor striking, then almost all instances of the kind must have the feature for the generic to be true (e.g., "Barns are red"). Many of the key factors in this conceptual approach are compatible with our view of prevalence as a predictive probability: What is required is that the distinctions made by conceptual accounts influence predictive probability, which mediates linguistic effects. Indeed, one version of this mechanism is already suggested by Leslie (2008), citing findings by Rothbart, Fulero, Jensen, Howard, and Birrell (1978), that a speaker's perception of the prevalence of the feature can be altered by virtue of its dangerousness or distinctiveness.

Default generalizations. The conceptual view of generics argues that it is not desirable to define generics in the same terms as quantifiers (e.g., a truth-functional threshold on prevalence), because generics are more basic or more fundamental than quantifiers (Gelman, 2009; Leslie, 2008). For example, according to certain measurements, young children have a nearly adult-like understanding of generics far earlier than they do with quantified language (Gelman, Leslie, Was, & Koch, 2015). Similarly, adults who are told novel information about categories using quantified language (e.g., "Most spiders shed their skin") and later asked to recall that same information will tend to recall quantified information as generics (e.g., "Spiders shed their skin" but not vice versa (Leslie & Gelman, 2012). As a result, generics are thought of as conveying default generalizations, whereas quantified language expresses more sophisticated and specific generalizations. Rather than be contradictory, our model can be seen as formalizing a default generalization as Bayesian belief updating via a threshold function whose threshold value is contextually informed by knowledge about properties. Alternative formulations of such a default generalization mechanism could be proposed and quantitatively tested against our account.

Striking generics. Leslie (2008) posits a special mechanism to treat generics about striking properties (e.g., "Rottweilers maul children"), which seem to be acceptable even when the prevalence of the property is quite low (e.g., very few rottweilers maul children). Indeed, Cimpian, Brandone, and Gelman (2010) found participants willing to endorse generics about striking properties (e.g., "Lorches have dangerous feathers") more so than that of neutral properties (e.g., "Lorches have purple feathers") at low levels of prevalence (e.g., when only 30% of the category had the property). Leslie (2008)'s argument for accepting striking generics is straightforward: These are relevant properties to know about if you want to survive. Our theory posits two mechanisms by which striking properties could influence generic endorsement: the prevalence prior and the referent prevalence.

The first observation is that striking properties are relatively rare in the environment: Most animals do not maul children, eat swimmers, or carry malaria. Thus, the prevalence prior distributions of these features may be hard to distinguish from properties that are

[9] We use the terms statistical and conceptual to refer to what Carlson (1995) referred to as "inductive" and "rules and regulations" views, respectively.

[10] This positive vs. negative counterinstance distinction is a rather technical consideration (though not without psychological foundation, see Leslie, 2008, p. 36) used to account for the difference between examples like "Robins lay eggs" vs. "Robins are female." For "Robins are female," male robins are positive counterinstances because they have an alternative property (being male). For "Robins lay eggs," male robins are negative counterinstances because they simply lack the property (i.e., male robins do not reproduce by some other method). We will not discuss this consideration any further other than to note that that the difference between "Robins lay eggs" and "Robins are female" can be explained by a different mechanism (e.g., informativity with respect to a prevalence prior). The notion that alternative features may come into play in generic interpretation may also be addressed with prevalence priors constructed with respect to the feature (described in the General Discussion).

just generally rare and distinctive, as we have assumed in our worked example of "Mosquitos carry malaria." Cimpian et al. (2010)'s experiments did not measure the prevalence prior distribution, but we have found in pilot work that the prevalence prior distribution changes to resemble that of a distinctive property when participants are supplied information about the dangerousness of a property.

The second mechanism by which strikingness could influence generic endorsement in our model is by speakers having a distorted perspective of how prevalent these features are within the referent category. That is, striking properties may be projected more strongly (i.e., higher predictive probability that future instances will have the property) than neutral properties.[11] An interesting finding was that this enhanced projectibility holds for both dangerous properties (e.g., people doing criminal actions) as well as neutral, distinctive properties (e.g., people taller than 6'5" Rothbart et al., 1978, Experiments 2 and 3). Thus, there is evidence for the influence of strikingness on predictive probability, which we posit is an intermediate representation between conceptual knowledge and generic endorsements. It is beyond the scope of this article to directly attempt to empirically distinguish these two potential mechanisms, though our experiments use a number of striking properties for which we measure the prevalence prior and referent prevalence.

Principled and statistical connections. Leslie (2008)'s construct of characteristic properties is similar to Prasada and Dillingham (2006)'s notion of a *k-property* (*k* for kind), a property that bears a *principled connection* to the kind. A property bears a principled connection to the kind if a (generic) sentence that appeals to the kind to explain the existence of the property makes sense (e.g., "Dogs, by virtue of being the kind of things that they are, are four-legged"). When such sentences are not endorsed (e.g., "Barns, by virtue of being the kind of things that they are, are red") while the simple generic (e.g., "Barns are red") is endorsed, then the property is thought to bear a *statistical connection* to the kind (a so-called *t-property*, for emphasizing that it is the tokens having the property that matters for generic endorsement). In addition to the difference in endorsements for *by virtue of* statements, Prasada et al. (2013) showed that different kinds of generic statements (e.g., striking properties, characteristic properties with low prevalence) support different kinds of inferences (e.g., *normative*: "Dogs are supposed to have four legs," *aspect*: "Having four legs is one aspect of being a dog"). Additionally, there is some evidence that the impact of principled connections on interpretations is separate from their influence on beliefs about prevalence (Prasada & Dillingham, 2006).

Both Prasada and Dillingham (2006) and Leslie (2008) surmise that a property may qualify as characteristic (or having a principled connection) if there exists an overhypothesis about that property for a relevant superordinate category (e.g., each kind of animal has a means for self-locomotion; Goodman, 1955; Shipley, 1993). Overhypotheses are naturally formalized as hierarchical Bayesian models, wherein a learner acquires knowledge at multiple levels of abstraction (e.g., learning from the same event about a particular dog, dogs in general, and animals in general; Kemp et al., 2007). These Bayesian heirarchical models yield differences in predictive probability that are particularly robust. Integrating a hierarchical model of kinds and properties with our model of generic language is a natural direction to understanding the computational underpinnings of conceptual relations and generics. Different conceptual structures may give rise to roughly the same distributions on prevalence and, thus, have similar generic endorsement profiles according to our model. Yet, the inferences that can be drawn from hearing these generics may differ, depending on their interactions with the putative hierarchical knowledge people bring to bear (a la the differences observed for "prevalence-matched items" in Prasada & Dillingham, 2006). Building hierarchical models of kinds and properties is a major undertaking in its own regard, but our formalism provides a way of connecting such a model with generic language.

Statistical Accounts

Relative and absolute generics. Our underspecified threshold model has clear antecedents in other statistical accounts, most notably Cohen (1999)'s theory of generics as a frequency adverb (e.g., "generally"). Cohen treats generics as a class comprising two qualitatively different types: *relative* and *absolute* generics. *Absolute generics* use a fixed, 50% threshold on prevalence: $p > .5$. That is, if a particular instance is more likely than not to have the feature, then an absolute generic is true. By contrast, *relative generics* are true based on a comparison to an alternative set of kinds $Alt(K)$, analogous to the categories that comprise the prevalence prior (the so-called *comparison class*, which we discuss further in the General Discussion). "Mosquitos carry malaria" is true because an arbitrary mosquito is more likely than an arbitrary member of an alternative kind to have the feature.

In our model, we treat all generics as relative. Attested differences in endorsement, then, emerge through the interplay of prior knowledge with our uncertain semantics.[12] Further, though Cohen's theory is framed in terms of probabilities, it is a fully deterministic, fixed-threshold account that only makes qualitative predictions about what is true and what is false (i.e., it has a deterministic semantics). In contrast, we propose a fully probabilistic semantics embedded within a Bayesian model that describes how context resolves the uncertain threshold; our theory is a joint semantic–pragmatic theory.

Cohen's and other statistical theories use a mechanism (not currently required in our account) known as *domain restriction* to explain the context-sensitivity of generics: contextually restricting the entities that go into the computation of prevalence (i.e., which robins do we look at to compute the probability of laying eggs among robins?). Cohen (1999) posits that prevalence is calculated by only considering entities that *could have some feature* in a contextually specified alternative set of features $Alt(F)$. For example, the property *lays eggs* induces a set of alternatives that are associated with modes of reproduction (e.g., *gives birth to live young*, *undergoes mitosis* . . .). "Robins lay eggs" (an absolute generic for Cohen, 1999) is evaluated by only considering female members of kinds, because only female members can plausibly satisfy one of the other reproductive property alternatives (i.e., the alternative features in $Alt(F)$). The inferential machinery behind

[11] This observation is also made by Leslie (2008, p. 42), who uses it as motivation for elevating striking properties to their special status in her theory.

[12] For a related theoretical argument against the "relative"/"absolute" distinction for gradable adjectives, see Lassiter and Goodman (2015).

domain restriction—how to determine $Alt(F)$—relies upon conceptual knowledge, but the details remain obscure (Carlson, 1995).[13] Our uncertain threshold model can be seen as one particular mechanism by which the domain may be restricted: The structure of the prior distribution over the prevalence of *lays eggs* is a reflection of an intuitive theory of reproduction (i.e., that only females lay eggs) and the uncertain threshold model uses that background knowledge to derive a property-specific interpretation. There may exist a refactorization of the uncertain threshold model to a fixed threshold where the listener has uncertainty about the relevant domain of restriction.

Generic as indexical. Sterken (2015) develops a novel analysis taking the context-sensitivity of generics as primary. This analysis draws analogy to other, inherently context-sensitive linguistic expressions: *indexicals* (e.g., "this" or "I"). Sterken (2015) uses this analogy to motivate a context-sensitive *quantificational force* as well as mechanism of domain-restriction (of the kind used by Cohen, 1999 and others). Our uncertain threshold semantics can be seen be a particular formalization of Sterken (2015)'s context-sensitive quantificational force. We have not had need for using domain restriction on the categories, though as noted above, it is a potential avenue for future development.

"Normal" accounts. A popular alternative view under the statistical banner draws on the intuition that generics often express something normative in the world (Asher & Morreau, 1995; Nickel, 2008, 2016; Pelletier & Asher, 1997). "Dogs have four legs" is then a good generic not because all dogs have four legs (regrettably, all do not) but were the world to function normally (e.g., dogs would not be involved in freak tractor accidents or be born with strange genetic mutations), then all dogs would have four legs. The idea that our beliefs about what is normal in the world influences our judgments about generalizations has intuitive appeal for rejecting accidentally true generics (e.g., "Supreme Court justices have even social security numbers") and resisting stereotyped language (e.g., "Boys are good at math"). Our theory does not directly formalize what is normal, though we argue that a speaker's beliefs about what is probable (that may relate to what is normal; see Icard, Kominsky, & Knobe, 2017) plays a role in endorsing and interpreting generalizations.

Underquantification. A proposal similar to our account concerning underspecification of generics has been made in the computational linguistics literature (so-called *underquantification*; Herbelot & Copestake, 2011). In their model, generics express an explicit quantified relation, specifically either "Some," "Most," or "All." This proposal is used to construct a set of features that accurately predicts (relative to human judgments) the quantified relationship expressed by the generic (analogous to a quantifier version of the *implied prevalence* task used in Cimpian et al., 2010; Gelman & Raman, 2003). Our semantic theory can be seen as a generalization of *underquantification* to a continuous interval of possible meanings. This distinction is relevant for the acquisition of the language of generalizations; we do not take as primary the quantified relations (e.g., "Some," "All"). Additionally, by using an underspecified threshold on a scale of probability, our formulation naturally extends to other scales and other kinds of generalizations (e.g., habitual language), where quantified relations like "Most" or "All" are not directly applicable.

Cue validity. By encoding knowledge about other categories, the prevalence prior distributions in the uncertain threshold model are deeply connected to the construct of *cue validity*, or the probability of the kind given the feature: $P(x \in k \mid x \in f)$(e.g., one's predictions about whether or not an entity is a mosquito, upon learning that it carries malaria). Cue validity is believed to play a role in understanding generic language (Khemlani, Leslie, & Glucksberg, 2012; Leslie, 2007; Prasada et al., 2013), but the details remain underspecified. The strongest view of cue validity is that it operationalizes an alternative hypothesis about the meaning of generic statements: "Mosquitos carry malaria" means "It is mosquitos that carry malaria." Indeed, empirically elicited cue validity has been shown to be highly correlated with endorsements of generics (Khemlani et al., 2012).

Cue validity is inverse prevalence; the two are related via Bayes' Rule: $P(k \mid f) = \frac{P(f \mid k) \cdot P(k)}{\Sigma_{k' \in K} P(f \mid k') \cdot P(k')}$. Knowledge about other categories k' enters in the denominator to compute cue validity for prevalence. Indeed this normalizing constant (the denominator) is equal to the expected value (i.e., the mean) of the prevalence prior distribution: $\mathbb{E}[P(p)]$. Thus, cue validity comes from a point estimate of the prevalence prior distribution, and information about cue validity can be derived from the constructs posited in our model. We return to the implications of this relationship in the General Discussion. For a more detailed, mathematical derivation of the relationship between cue validity and prevalence priors, see Appendix A.

Baseline Models for Quantitative Comparisons

In our empirical studies below, we compare our model to three alternatives. These alternative models do not represent any of the extant theories of generics described above; no extant theory is sufficiently precise to yield quantitative predictions. Instead, these models are designed to interrogate the theoretically substantive components of our model. There are three such components: (1) property knowledge in the form of a prior distribution over prevalence $P(p)$, (2) endorsement as a decision-theoretic process of uttering the generalization versus not uttering it, and (3) vagueness in the semantics of a generalization (i.e., an uncertain threshold). In our empirical studies, we compare the uncertain threshold model to an alternative, lesioned model that lacks the vagueness in meaning, assigning a fixed semantics to the generalization (i.e., analogous to a quantified statement) but that has the same prevalence prior and the decision-theoretic architecture. There are no correspondingly simple ways to lesion the other two components (property knowledge or speaker decision) that still produces quantitative, context-sensitive predictions. Instead, we compare our model to two regression models based on empirically elicited referent prevalence and cue validity, which provides some interrogation of the necessity of property knowledge in the form of a full distribution on prevalence. In addition to serving as alternatives, these baseline regression models help us understand the statistical properties of our experimental materials and provide a comparison with standard methods in the psychological literature on generics (e.g., Khemlani et al., 2012).

[13] However, see Cohen (2004) for a discussion of how his semantic constraints relate to different kinds of generics and different kinds of conceptual representational frameworks used in cognitive science.

Interim Summary and Overview of Experiments

We have introduced the first quantitative theory of the language of generalization and discussed the relationship of this account to extant theories of generics. Above, we presented simulations showing how the model predicts endorsements for classically puzzling generic statements. These predictions depended upon the background knowledge about the property $P(p)$ as well as the referent prevalence p believed to be true for the category. For each of these we chose intuitive values for the parameters of the model (i.e., the prevalence priors and referent prevalence levels). In what follows, we test this theory empirically for a wide range of generalizations, including generalizations of different types (categories, events, and causes). We do this by both measuring and manipulating background knowledge and referent prevalence and predicting human endorsements of generalizations.

In Case Study 1, we examine generalizations about categories expressed in generic language. We first measure endorsements for 30 generic statements about familiar categories, revealing an entire continuum of endorsements (Experiment 1a). We then measure the corresponding prevalence priors and referent prevalence using a prevalence prior elicitation task (Experiment 1b). We compare the quantitative fits of our model to the three alternative models described above. This Case Study is an empirical version of several of the worked examples in the section above.

In Case Study 2, we examine generalizations about events expressed in habitual language while manipulating the referent prevalence. We measure prevalence priors for events of people doing various actions (e.g., people running, hiking, climbing mountains; Experiment 2a). We then measure endorsements of habituals for statements about novel actors (e.g., "John runs") given referent prevalence information (e.g., "In the last two months, John ran three times."; Experiment 2b). Finally, we test whether referent prevalence in our model is best thought of as a past frequency (e.g., the number of times John has run in the past) or a prediction about the future (e.g., the number of times a speaker expects John to run in the future) by experimentally manipulating future predictions while keeping constant past frequency (Experiment 2c). We answer this question by testing the quantitative fits of two versions of our model: one in which the speaker is conveying past frequency and one in which the speaker is conveying their predictions about the future.

Case Study 3 experimentally manipulates the prevalence prior in the domain of causal language (e.g., "Herb X makes animals sleepy"). Experiment 3a measures the (manipulated) prevalence prior, confirming that our manipulation influenced participants' beliefs about the prevalence of the feature across different categories. Experiment 3b measures the corresponding influence on causal endorsement, finding an effect of the manipulated background knowledge in the way predicted by the uncertain threshold model.

Case Study 1: Generic Language

Learning from generic language (i.e., generalizations about categories; e.g., "Dogs bark") is believed to play a central role in concept and theory formation (e.g., Gelman, 2004), stereotype propagation (Rhodes et al., 2012), motivation (Cimpian et al., 2007), and many other facets of everyday reasoning. In addition, generics have been the case study of choice for the semantics of the language of generalization because of their tantalizing similarity to quantified statements (e.g., "Most dogs bark"). However, intuitions and empirical data argue that generics simply do not reduce to quantified statements in a simple way (e.g., Cimpian et al., 2010; Khemlani et al., 2012; Prasada et al., 2013).

We first investigate how the uncertain threshold endorsement model predicts actual human endorsements of generic statements. We measure endorsement for 30 generic sentences that cover a range of conceptual distinctions previously discussed in the empirical literature on generics (Prasada et al., 2013): characteristic features displayed by a majority (e.g., "Ducks have wings"), characteristic features displayed by a minority (e.g., "Robins lay eggs"), features that are striking or dangerous (e.g., "Mosquitos carry malaria"), noncharacteristic features displayed by a minority (e.g., "Robins are female"), and features that are totally absent (e.g., "Lions lay eggs"). We further craft sentences with the goal of eliciting the full range of acceptability judgments (intuitively: true, false, and indeterminate) for generics with properties of low, medium, and high referent-prevalence (Experiment 1a). We examine generics about animal categories to reliably measure the prior belief distribution over the prevalence of features $P(p)$. The prevalence elicitation procedure (Experiment 1b) includes measurements of the referent-prevalence p for different categories (e.g., $P(x \text{ lays eggs} \mid x \text{ is a robin})$), allowing us to generate predictions for the endorsement model (Eq. 3) as well as for simpler, alternative models.

Experiment 1a: Generic Endorsements

In this experiment, we elicit human endorsements for generic sentences taken from the linguistic and psychological literatures (Prasada et al., 2013). The goal of this study is to elicit high variability of endorsements for generic statements about animal categories.

Method.

Participants. We recruited 100 participants over Amazon's crowd-sourcing platform Mechanical Turk (MTurk). Participants were restricted to those with U.S. IP addresses and with at least a 95% MTurk work approval rating (the same criteria apply to all experiments reported). Four participants were excluded for failing to recall the button corresponding to agreement in the forced-choice task. Five participants self-reported a native language other than English; removing their data has no effect on the results reported. The experiment took about 3 min and participants were compensated $0.35.

Procedure and materials. Participants were shown 30 generic sentences in a randomized order. They were asked to press one of two buttons (P or Q; randomized between-participants) to indicate whether they agreed or disagreed with the sentence (see Figure 6A for the full list). The 30 sentences covered a range of conceptual categories described above. Approximately 10 true, 10 false, and 10 uncertain a priori truth-value generics were selected. As an attention check, participants were asked at the end of the trials which button corresponded to "Agree." Four participants were excluded for failing this trial.

Results. As a manipulation check, the first author assigned an *a priori* truth judgment (true/false/indeterminate) to each stimulus item. As one would expect, there were substantial differences in empirical endorsements: true generics were almost universally endorsed (Maximum A-Posteriori estimate and 95% Bayesian credible interval (CI) of endorsement probability: 0.93 [0.91, 0.94]); indeterminate generics were endorsed at a rate *less* likely than chance (0.38 [0.35, 0.42]) but substantially more than false generics (0.08 [0.06, 0.09]).

Ideally, a complete theory of genericity should be able to explain statements that are endorsed completely, rejected completely, and the gradedness between the extremes. We observe gradedness among our 30 examples covering a continuum of endorsement values (Figure 6A). Such a continuum of judgments is already evidence against any theory that only predicts categorically whether a generic statement is true or false. We next measure the prevalence prior distributions and use them to articulate a set of quantitative models that try to predict this quantitative variability in endorsements.

Experiment 1b: Prevalence Prior Elicitation

The prevalence prior $P(p)$ in Eq. 1 describes the belief distribution on the probability of a given feature (e.g., lays eggs) across relevant categories. To get an intuition for the kind of knowledge encoded in this belief distribution, imagine you are walking outside and come across an instance of your favorite kind of animal (e.g., a reindeer). What is the chance it is female? Your answer will probably depend upon the percentage of the category that you believes to be female (e.g., the percentage of female reindeer, approximately 50%). What is the chance that it lay eggs? Again, this depends upon the percentage of the category that you believe *lays eggs*, which then further depends on the particular kind of animal under consideration: If you are thinking of a reindeer, the answer is probably 0%; if you're thinking of a peregrine falcon, the probability is similar to the *being female* probability (50%) because female peregrine falcons lay eggs. That is, the answer to how many of an arbitrary category is likely to lay eggs is either roughly 50 or 0%, depending on the kind of creature you may bring to mind.[14]

The thought experiment decomposes the prevalence prior $P(p)$ into a prior distribution on kinds $P(k)$ and then a conditional probability of the prevalence given the kind $P(p \mid k)$. This decomposition can be used to measure the prevalence prior for familiar properties $P(p) = \int_k P(p \mid k) dk$ as a stand-in for a richer intuitive theory that could give rise to prevalence judgments. We measure prior distributions empirically for the set of properties (e.g., *lays eggs*, *carries malaria*; 21 in total) used in our generic sentences from Experiment 1a. To create a larger set of properties, we reverse-code responses for five properties to create their corresponding negative properties (e.g., we create a property "doesn't have beautiful feathers" by subtracting from 100% the responses for "has beautiful feathers").[15]

Method.

Participants. We recruited 60 participants over Amazon MTurk. Three participants were accidentally allowed to complete the experiment for a second time, so we excluded their second responses (resulting in $n = 57$). Two 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 min and participants were compensated $1.00.

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 (see Figure 4). Participants first were shown six animal categories randomly sampled from a set corresponding to referent-categories of the generic sentences used in Experiment 1a (e.g., robins, mosquitos) and were asked to generate five animal kinds of their own (Figure 4A). A column then appeared to the right of the animal names with a property label in the column header (e.g., *lays eggs*). Participants were asked to fill in each cell with the percentage of members of each of the species that had the property (e.g., "50%"; Figure 4B). Eight property columns in total appeared in the table. This whole procedure was repeated two times (two trials). In total, each participant generated 10 animal names and reported on the prevalence of 16 properties for 22 animals (their own 10 and the experimentally supplied 12).

Qualitative results. The elicited prior distributions have a diversity of shapes (eight examples shown in Figure 5A) that are qualitatively consistent with the schematic prior distributions used in the Worked Examples section (Figure 2A). The property *being female* is present in almost all categories in almost exactly the same proportion, whereas priors for properties such as *laying eggs* or *having spots* exhibit more structure represented by the multimodality of these distributions. *Being red* exists mostly at extremely low prevalence levels (i.e., 0 prevalence) but also at high prevalence levels (e.g., 50 or 100%), whereas *carrying malaria* is really only present at low prevalence levels. This diversity is relevant because our endorsement model makes different predictions depending on the shape of these distributions.

Modeling the prevalence priors. To incorporate the uncertainty in our measurement of the prevalence prior into the endorsement model's predictions, we build a Bayesian statistical model of the prior elicitation data. We approximate the prevalence distribution for each property (e.g., *lay eggs*) with a Mixture of Betas model, which assumes that the data generated for each kind comes from one of two underlying Beta distributions.[16] We specify one of these distributions *a priori* to represent kinds of animals who do not have a stable causal mechanism that could give rise to the property (e.g., lions and lay eggs), which results in prevalence or prevalence values close to or equal to 0. This *null distribution* is potentially present for all features and acts in exactly the same way (i.e., the lack of producing the feature).[17] The second distribution represents kinds of animals who could have such a mechanism, and the two parameters of this distribution are not specified *a priori* and are not the same for all properties, but are inferred on a property-wise basis from participants' responses. The *Mixture of Betas* distribution has a third free parameter (for each property):

[14] The subjective probability may, in fact, be nonzero and small as opposed to 0. Nonzero probabilities allow for the intuitive possibility of a reindeer that, by some terribly improbable set of circumstances such as a genetic mutation, lays eggs.

[15] This reverse-coding assumes for these properties that logical negation tokenizes its own threshold instead of being derived compositionally via bivalent logical negation. For related empirical and modeling investigations regarding resolving compositional versus noncompositional negation in the context of gradable adjectives, e.g., "not happy" vs. "unhappy", see Tessler and Franke (2018).

[16] The Beta distribution is chosen because the support of this distribution is numerical values between 0–1, exactly the form of the response data in the prior elicitation task.

[17] This assumption is similar in spirit to that used by *Hurdle Models* of epidemiological data, where the observed count of zeros is often substantially greater than one would expect from standard models, such as the Poisson (e.g., when modeling adverse reactions to vaccines; Rose, Martin, Wannemuehler, & Plikaytis, 2006)

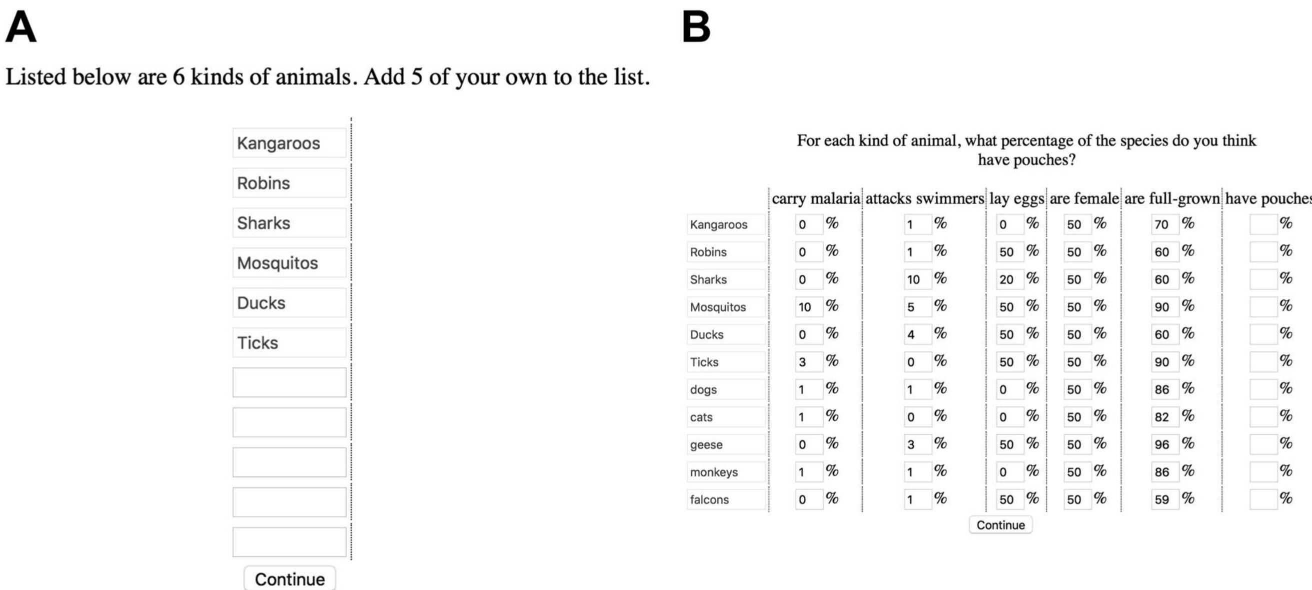

Figure 4. Prior elicitation task. (A) Participants first generated animal names after seeing six example categories. (B) One feature at a time, participants estimated the percentage of the category with the feature, for each category.

the relative contribution of the null distribution (e.g., we expect the null distribution to contribute not at all to properties like *being female*, for which almost all categories have at least some members with the property). To ensure the *Mixture of Betas* model of the prior is not overly complex, we fit an additional model that represents only a single underlying distribution (*Single Beta*) for a comparison. For more details about model implementation and inference, see Appendix C.

The prior distributions over prevalence are well modeled as a mixture of two Beta distributions and not as a single Beta distribution (Figure 5B; red vs. blue lines). The *Single Beta* model provides a good fit for the *being female* distribution, but overly smooths the other distributions, washing out the latent structure in participants' responses. One property that the *mixture of Betas* model does not perfectly capture is the prior distribution over the feature *lays eggs*. The empirical distribution is tri-modal, with reliable modes at 0, 50, and 100%; a simple two-component mixture model has no way to account for such a tri-modal distribution.[18] A more complex model (i.e., one with three mixture components) would be necessary to perfectly account for this item. Using a three-component model for this distribution does not change the resulting model predictions and we maintain the simpler two-component mixture for uniformity.

Endorsement Model Comparison

We can now compute the predictions of our endorsement model. We describe first the behavior of a set of alternatives models that have been previously proposed in the literature and an alternative form of our proposed endorsement model before proceeding to the results of the uncertain threshold endorsement model.

Baseline models. We present two baseline, quantitative models that have previously been used in the empirical literature on generic language. In addition to serving as alternatives, these regression models help us understand the statistical properties of our experimental items. First, we estimate how well referent prevalence itself predicts generic endorsement (e.g., does the fraction of robins that lay eggs predict the felicity of "Robins lay eggs"?). Second, we include cue validity—the probability of a kind given the feature—as a second predictor in a linear model. We fit these models using standard maximum-likelihood techniques and model uncertainty in the input measurements (i.e., referent prevalence, cue validity) by bootstrapping those data (for a justification, see *Supplementary Model Criticism* in Appendix C).

Referent prevalence. From the prevalence prior data (Experiment 1b), we estimate participants' beliefs about the referent prevalence (e.g., the percentage of robins that lay eggs) and use it to predict endorsement. We find a little over half of the variance in the endorsement data are explained this way ($r^2(30) = 0.51$; $MSE = 0.08$; Figure 5B upper-left facet). Referent prevalence alone predicts a fair amount of variance because our stimulus set includes generics that are true with high prevalence properties (e.g., "Leopards have spots") and false with low prevalence properties (e.g., "Leopards have wings").

[18] The third mode at 100% is not entirely attributable to categories for which all members could be female (e.g., chickens). Instead, it appears that some participants are responding that 100% of several different kinds of birds (e.g., robins) lay eggs. This may result from participants implicitly only considering female members of the category as relevant to answer a question about a reproductive capacity like *lays eggs*; this restriction of what enters into the prevalence computation is known as *domain restriction*, is posited in several theories of generics (e.g., Cohen, 1999), and has been observed in other prevalence elicitation tasks (e.g., Prasada et al., 2013).

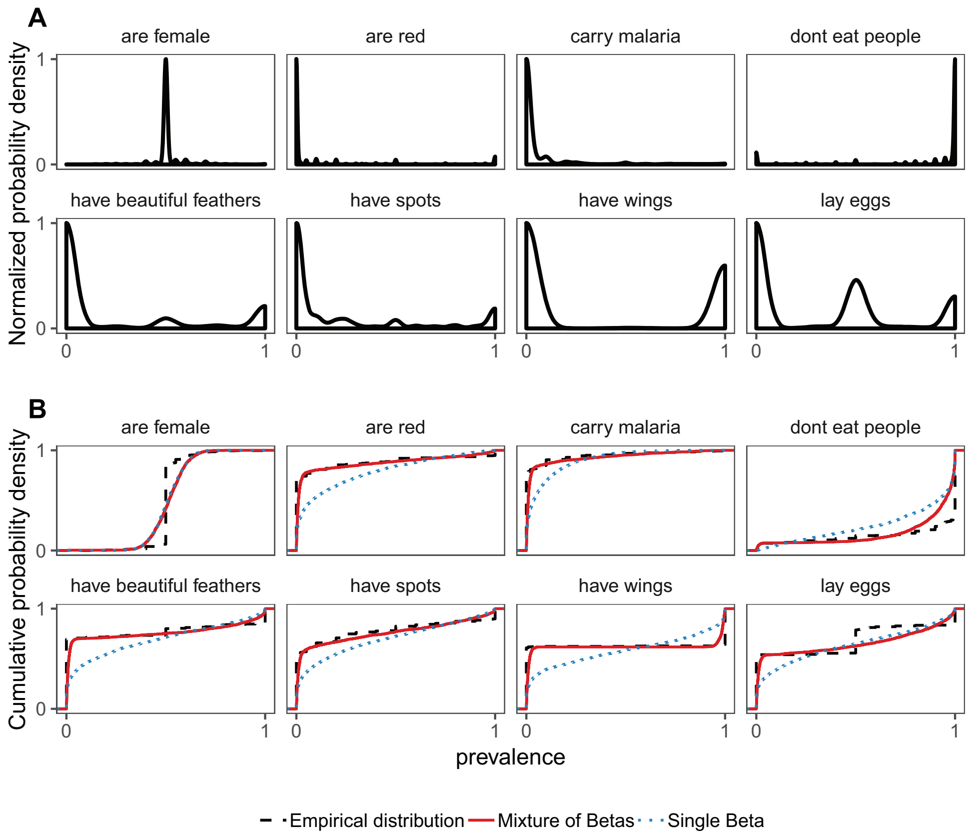

Figure 5. (A) Empirically elicited prior distributions over prevalence for eight properties. (B) Cumulative density plots reveal that a model of a mixture of two Beta distributions does substantially better at capturing the structure of the priors than a single Beta distribution. Distributions are the posterior predictive distributions for the models of the prior and the raw empirical distribution. A completely uniform distribution would be represented as the $y = x$ line. See the online article for the color version of this figure.

Large deviations from an account based purely on referent prevalence remain: Generics in which the referent-category has intermediate prevalence (prevalence quartiles 2 and 3: 16% < prevalence < 64%), are not at all explained by referent prevalence ($r^2_{Q2,3}(15) = 0.01$; $MSE = 0.14$). This includes generics that are judged true with relatively low referent prevalence (e.g., "Mosquitos carry malaria") and false with relatively high referent prevalence (e.g., "Sharks don't eat people").

Cue validity and referent prevalence. Cue validity indexes the diagnosticity of the feature for a kind, given formally by the probability of a kind given the feature $P(k \mid f)$. As discussed in the *Relationship to Other Semantic Theories* section above, cue validity is linearly related to expected value of the prevalence prior distribution (see Appendix A for derivation). Cue validity acts using a point estimate of the prevalence prior distribution, a single metric that summarizes the prior distribution.

Though the cue validity of a property for a category can be derived from the prevalence prior distribution, previous empirical studies of generics have estimated cue validity from different empirical sources (Cree, McNorgan, & McRae, 2006; Khemlani et al., 2012). In the empirical literature on generics, researchers often ask directly about the cue validity probability (e.g., "There is an animal that lays eggs. What is the probability that it is a robin?"; Khemlani et al., 2012), though in the broader literature on semantic memory a *free production* paradigm is often used (e.g., "X lays eggs. What do you think X is?"; Cree et al., 2006). We found that these two ways of estimating cue validity diverge for a number of key cases (see Appendix B). Most notably, undiagnostic features (e.g., *is female*), which in theory have a cue validity close to zero, were rated as having intermediate cue validity (around 0.5) in the *direct question* paradigm, but not in the *free production* paradigm. In response to a direct question such as "There is an animal that is female. What is probability that it is a robin?", participants seem to want to respond "I don't know" by placing the slider bar at the midpoint, rather than reporting their intuitive base rate that a random animal would be a robin. The free production paradigm did not produce artifacts such as this one, and we chose to use it as the more veridical estimate of cue validity. For a detailed analysis of the different cue validity measurements and comparison with cue validity derived from the prevalence prior, see Appendix B.

A linear model that uses predictors for both referent-prevalence and cue validity does a better job at explaining the endorsement data than just prevalence alone ($r^2(30) = 0.73$; $MSE = 0.04$). This model is able to account for the endorsements of examples like "Mosquitos carry malaria" (model endorsement and bootstrapped

95% CI 0.85 [0.72, 0.86]) and "Lions have manes" (0.79 [0.61, 0.84]), as these features are very diagnostic of the kind (generic endorsement both >0.9). Deviations, however, still remain. For example, "Robins lay eggs" still receives only intermediate endorsement by this model (0.68 [0.56, 0.69]; human endorsement = 0.94 [0.87, 0.97]), and "Mosquitos don't carry malaria" is misjudged to be a pretty good statement (0.58 [0.41, 0.58]; human endorsement = 0.07 [0.04, 0.14]).

"Robins lay eggs" and "Mosquitos don't carry malaria" highlight a shortcoming of reducing structured prevalence prior distributions to single point-estimates of cue validity. *Lays eggs* is a somewhat diagnostic feature for birds, but there are many kinds of birds, and the feature is not itself diagnostic for a particular kind of bird like robins. Thus, the cue validity is low even though robins are in the distinctive part of the *lays eggs* prevalence prior distribution (Figure 2A bottom). Furthermore, cue validity cannot distinguish *undiagnostic* features (features present in almost every category; e.g., *not carrying malaria*) from *false* features (features that are absent a particular category; e.g., *lions* and *lay eggs*; see Appendix B for more discussion of this distinction); the cue validity can be near-zero for different reasons. Such a model makes the wrong prediction for nondistinctive properties with high referent prevalence (e.g., "Mosquitos don't carry malaria").

Communicative endorsement models. Our underspecified-threshold model considers how well the generalization would bring a naive interpreter's prior distribution on prevalence $P(p)$ (Eq. 1; e.g., the prevalence of laying eggs among other animals) in line with the speaker's belief about the referent prevalence (p in Eq. 3; e.g., the prevalence of laying eggs among robins). As described in the *Baseline Models for Quantitative Comparisons*, there are several substantive components to this hypothesis: (a) property knowledge in the form of a prior distribution over prevalence $P(p)$, (b) endorsement as a decision-theoretic process of uttering the generalization versus not uttering it, and (c) vagueness in the semantics of a generalization. We construct an alternative, lesioned model by removing the vagueness in meaning, assigning a fixed semantics to the generalization (i.e., analogous to the quantified statements) but keeping the prior and the decision-theoretic architecture in place. There are no correspondingly simple ways to lesion the other two components (context or speaker decision) while still producing a model that makes quantitative, context-sensitive predictions.

For both the fixed-threshold and full uncertain-threshold endorsement models, we build joint Bayesian data analysis models of the referent prevalence p, prevalence priors $P(p)$ (both from Experiment 1b data), and the endorsement data (Experiment 1a). Predicting the data from both Experiment 1a and 1b by a single, joint-inference model makes our assumptions explicit about how these data were generated and is the proper way to represent the uncertainty in our measurement of the prior elicitation data (see Appendix C and Figure C1 for further model specification details). Empirically elicited referent prevalence and prevalence prior data (Experiment 1b) directly constrain the parameters that generate those quantities in the model (p in Eq. 3 and $P(p)$ in Eq. 1, respectively). The prevalence prior $P(p)$ is modeled as a mixture of Betas, and referent prevalence is modeled by a single Beta distribution. The endorsement data (Experiment 1a) is modeled by our endorsement model (Eq. 3), which has one free parameter λ. To learn about the credible values of the parameters of the joint-inference model and resulting model predictions, we ran an incrementalized version of MCMC (Ritchie, Stuhlmüller, & Goodman, 2016) for three chains of 150,000 iterations, discarding the first 50,000 for burn-in.

Lesioned model (no vagueness). For a strong alternative model, we lesion the uncertain threshold model so that it has a fixed threshold θ. We test the strongest, possible fixed-threshold model by searching for the best possible single threshold that fits the data. However, a fixed-threshold model will have to accommodate responses that are literally inconsistent with the threshold (i.e., a participant endorsing a generic when the referent prevalence is less than θ); thus, we outfit this model with an additional extrinsic noise parameter, to allow for random guessing. Thus, the fixed-threshold alternative model claims that participants make an information-theoretic decision, taking into account the interpreter's prior distribution on prevalence $P(p)$, using a fixed-threshold semantics, where deviations from a pure information-theoretic decision are accounted for by noise. This alternative model has two additional parameters to our uncertain threshold model (the value of the fixed-threshold and the proportion of noise responses).

To evaluate the fixed-threshold model, we examine model predictions as well as the posterior distribution over latent parameters of the model (referent prevalence, prevalence priors, the optimality, fixed-threshold, and noise parameters) given the observed data. The Maximum A-Posteriori value and 95% highest probability density interval for the inferred (fixed) threshold and noise parameters are 0.34 [0.25, 0.37] and 0.20 [0.18, 0.22], respectively. The inferred optimality parameter in Eq. 3 is 0.34 [0.25, 0.37]. Figure 6B (bottom right subplot) shows the fixed-threshold model's ability to predict the generic endorsement data ($r^2(30) = 0.9329$; $MSE = 0.010964$). Though the model is able to capture a lot of the variance, it only makes three kinds of judgments: true, false, or neither, similar to purely semantic accounts of generics. It treats "Tigers eat people" (0.90 [0.89, 0.91]) as good a statement as "Peacocks have beautiful feathers" (0.90 [0.89, 0.91]), though participants give a substantially weaker endorsement of the former ("Tigers eat people" = 0.69 [0.59, 0.77]; "Peacocks have beautiful feathers" = 0.99 [0.94, 1]). Similarly, "Lions lay eggs" (0.10 [0.09, 0.11]) is judged to be just as bad as "Mosquitos attack swimmers" (0.10 [0.09, 0.11]), though participants rate the former as completely false (0.01 [0, 0.06]) while the latter is kind of true (0.38 [0.28, 0.48]). The fixed-threshold alternative model is unable to make these fine-grained distinctions because it uses the same semantic threshold in all contexts.

Uncertain threshold model. Our underspecified threshold model is the same as the fixed-threshold model, except that rather than having a fixed θ for all contexts, the model infers property-specific θ's. We use the same Bayesian data analysis approach, dropping the additional parameters required for the fixed-threshold model (the fixed-threshold and noise parameters). Thus, this model has two fewer parameters than the fixed-threshold model above.

We first examine the posterior predictive distribution on the prevalence prior and referent prevalence data to ensure that the joint-inference model does not distort these parameters at the service of predicting the endorsement data (e.g., such a distortion could manifest by the joint-inference model inferring that 100% of mosquitos carry malaria to predict that "Mosquitos carry malaria" is a good utterance). This is an important step in model validation because it reveals whether the model's predictions are derived

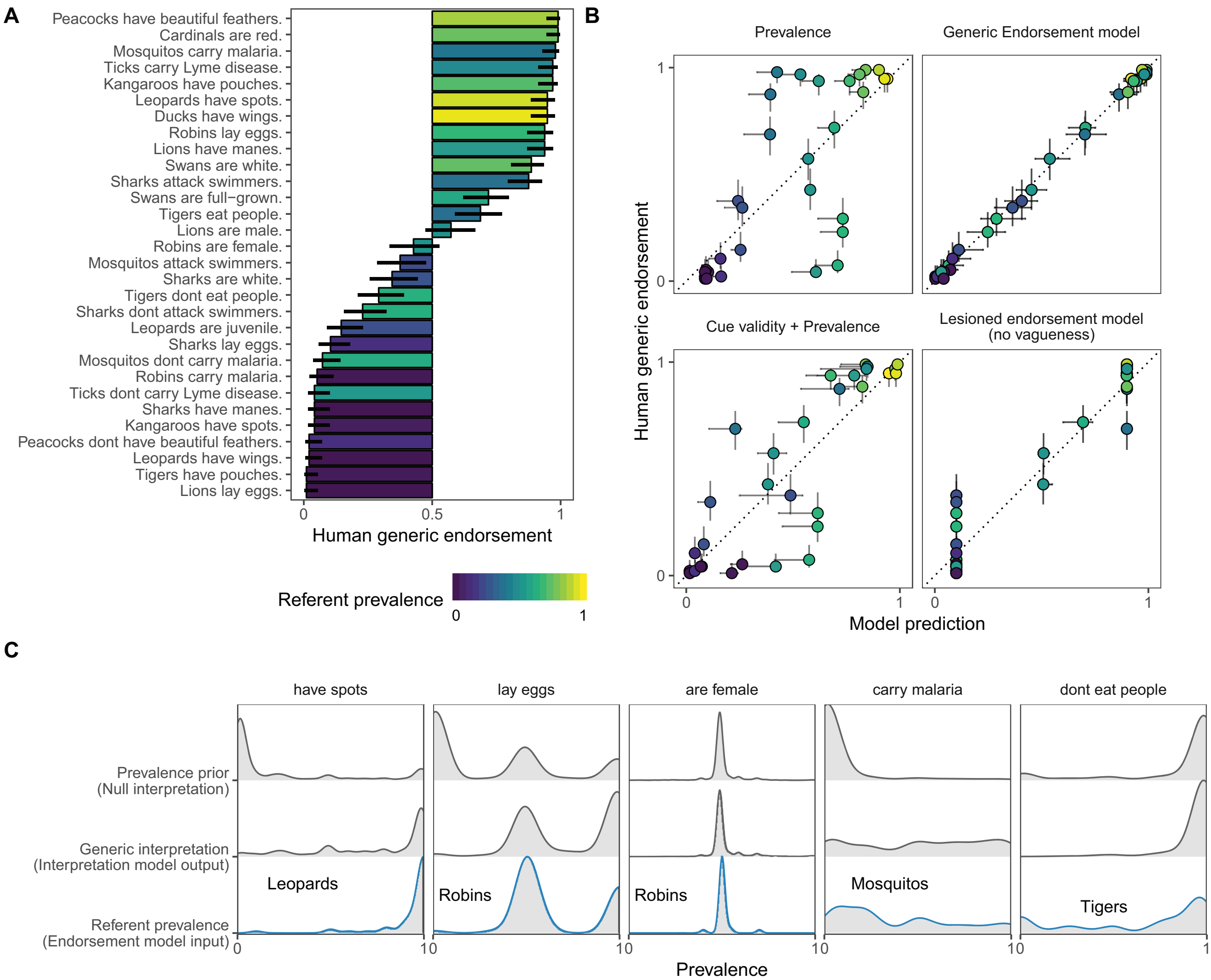

Figure 6. Endorsing generalizations about categories. (A) Human elicited endorsements for 30 generic sentences reveal a continuum of endorsements. (B) Model fits for the uncertain semantics speaker model (upper right), a fixed semantics speaker model (lower right), and regression models based on referent prevalence alone (upper left) and Prevalence + Cue validity (lower left). (C) Five example empirical prevalence priors, model-predicted generic interpretations, and empirical referent prevalence (speaker belief) distributions. See the online article for the color version of this figure.

from intuitively plausible values of the parameters (e.g., that not all mosquitos carry malaria). The joint-inference model captures the prior elicitation data (e.g., the probability of carry malaria among various species) and the referent prevalence data (e.g., the prevalence of carrying malaria among mosquitos) as well as it did when we analyzed these data in isolation (i.e., without conditioning on the endorsement data) ($r^2_{\text{prevalence prior parameters}}(60) = 0.9997$; $r^2_{\text{referent prevalence parameters}}(63) = 0.9989$; see Appendix D, Figure C2). This result confirms that the theoretically interesting predictions of this model—predictions of generic endorsement—are based on intuitively meaningful model components (i.e., the shapes of the prevalence distributions). Finally, the inferred optimality parameter in the endorsement model (Eq. 3) is 2.47 [2.18, 2.75], a range consistent with the literature on similar models.

As we see in Figure 6B (top right), the uncertain threshold endorsement model explains nearly all of the variance in human endorsements ($r^2(30) = 0.9978$; $MSE = 0.00035185$). Examining the relevant model components that give rise to these predictions further reveals the intuition for why the model makes the predictions that it does (Figure 6C). Recall the endorsement model's alternative utterance is a null or silent utterance and, thus, the prevalence prior distributions are exactly what the interpreter model would believe if the speaker does not produce the generic. The endorsement model then decides whether maintaining the listener's prior or updating it with the uncertain threshold semantics would better get the listener to guess the correct prevalence for the category (correct in the mind of the speaker). The model rates "Robins lay eggs" as a good utterance because the prevalence posterior implied by the generalization is similar to the referent prevalence.[19] On the other hand, the hypothetical interpretation for "Robins are female" is almost indistinguishable from the prevalence prior, because the prevalence prior has such low variance (almost all animals have female members in exactly the same proportion); the endorsement model then has no basis to prefer silence or the generic statement and the model predicts that the utterance should be neither good nor bad—endorsement around 0.5—also approximately the proportion of participants who endorse the statement. "Mosquitos carry malaria" is an interesting case because the prevalence prior has high variance (i.e., participants are highly uncertain about the prevalence of carrying malaria among categories). As a result, the generic interpretation also has high variance; still, the generic interpretation is more consistent with the referent prevalence than the prevalence prior, and the endorsement model predicts "Mosquitos carry malaria" is a good utterance. Finally, an utterance with high referent prevalence, such as "Tigers don't eat people," is predicted to have low endorsement because the generic would be misleading; even if most tigers do not eat people, saying "Tigers don't eat people" implies that all do not eat people, which is too strong.

Discussion

Generic language is the premier case study for generalizations in language. Generics have been studied extensively in the cognitive and developmental psychological literatures and have deep implications for wide ranging phenomena from stereotype propagation (Rhodes et al., 2012) to motivation (Cimpian et al., 2007). Heretofore, no models have been articulated with enough precision to make quantitative predictions about endorsement decisions, deciding whether a statement is true or false. This empirical case study demonstrates that a semantics based on the prevalence of the feature is tenable despite of alleged counterexamples (e.g., "Robins lay eggs" vs. "Robins are female"). The key theoretical insight is that the truth-functional semantics is underspecified, or vague, and resolved in context by a process of probabilistic inference. Our model provides a clear delineation of world knowledge (formalized as a prevalence prior) from the semantics of generics. We return to this point, and its implications for theory-building, in the general discussion.

In explaining the variable endorsements of generics, we related the referent prevalence (e.g., the percentage of robins that lay eggs) and the prevalence prior (e.g., the prevalence of laying eggs for different kinds of animals) to the endorsement of the generalization (e.g., "Robins lay eggs") via an information-theoretic communicative model where the meaning of a generic is simple but underspecified. In this case study, we used generic statements about familiar animal categories, which has long been the cleanest domain for testing semantic theories of generics by providing minimal comparison like "Robins lay eggs" versus "Robins are female." However, modeling familiar category generics is a correlational analysis: The relevant quantities in the model were measured rather than manipulated. We now seek to demonstrate how these quantities are causally related to endorsement, by manipulating referent prevalence (Case Study 2) and prevalence priors (Case Study 3). In addition, we take this opportunity to highlight the generality of the theory, by performing these additional empirical tests in different domains for generalization: events and causes.

Case Study 2: Habitual Language

As with instances of categories, particular events like "Mary smoked yesterday" can be generalized into *habitual sentences*: "Mary smokes." It is believed that an analysis of generics should lend itself naturally to be extended to an analysis of habituals (e.g., Carlson, 2005; Leslie, 2008), but no such analysis or empirical data has directly connected the two. In our second case study, we focus on habituals about people's behaviors that take the form: *singular noun phrase + present tense simple verb phrase* (e.g., "Mary smokes cigarettes"). Learning about the behaviors of others is useful because they tell us about what that person is like more generally (e.g., Repacholi & Gopnik, 1997; Seiver, Gopnik, & Goodman, 2013), and when children describe their lives to others, a surprisingly large amount of the language produced concerns the actions of people close to them articulated using habituals (e.g., "My brother works part-time at the restaurant"; McGuire & McGuire, 1986).

To test the generality of our theory, we use the same computational model and follow the same general experimental structure as in the first case study. We take the event analogue of prevalence to be the rate with which the event occurs (e.g., how often Mary smokes).[20] We test the model by first measuring the prevalence (rate) prior distribution for various actions (e.g., how often different people smoke cigarettes; Experiment 2a). We then measure endorsements of habitual statements while manipulating the referent prevalence (Experiment 2b), and use our computational model to predict habitual endorsements. By describing novel characters to participants, we are able to directly manipulate the referent prevalence, which we were unable to do for familiar categories in Case Study 1.

Finally, if habituals (and generics) are truly language for conveying generalizations, they should reflect speakers' expectations, not merely their observations. This intuition is sometimes expressed as an intensional meaning component of generics (Dahl, 1975). For example, imagine a very small town where by total coincidence, all residents chew sugarless gum. Endorsing the sentence "Residents of this town chew sugarless gum" seems to commit the speaker to believing that it is not sheer happenstance, but that there is some underlying cause that supports the counterfactual implication that were a new person to become a resident of the town, they too would likely chew sugarless gum.

Our computational model predicts endorsement rates given a referent prevalence p. Using this model, we can ask quantitatively how well prevalence conveyed by a speaker represents the actual, objective frequency in the world (e.g., the rate at

[19] We again note that the empirical prevalence prior for *lays eggs* is tri-model with peaks at 0, 50, and 100%. As a result, the generic interpretation is bi-modal: the listener is led to believe either 50 or 100% of robins lay eggs. An interesting find, the endorsement model still makes the correct prediction because the referent-prevalence for laying eggs among robins is also bi-modal (peaks at 50 and 100%). That is, domain-restriction (resulting in a response of 100%) seemingly occurs for some participants for *both* the prevalence prior and referent prevalence measurements, and this allows the model to predict correctly that "Robins lay eggs" is a good generic utterance.

[20] Specifically, for the generalization "Mary smokes," the instances being generalized are *instances of Mary*.

which a person has smoked cigarettes in the past) or a subjective, predictive belief in the head (e.g., the rate at which a person is expected to smoke in the future). Such a distinction would support the intuitions about sugarless gum residents and could explain why "Supreme Court Justices have even social security numbers" sounds strange even if nine out of the nine current justices have even social security numbers (Cohen, 1999): Our predictions about the evenness of the next justice's social security number are driven by strong prior beliefs that selection for the Supreme Court is uncorrelated from the numerical properties of one's social security number; the current observations are not enough in this case to sway those beliefs. We examine this aspect of the theory by measuring endorsements of habituals when causal forces intervene on the world (e.g., the person buys a pack of cigarettes; Experiment 2c) as well as participants' predictions about the likely frequency of the event in the future. We then compare habitual endorsement models based on speakers aiming to convey the objective, past frequency or their subjective, future expectation.

Experiment 2a: Measuring the Prevalence Prior for Events

To generate model predictions for habitual endorsements, we first elicit the prior distributions over rates for different events. For language about the behaviors of people, $P(p)$ represents a language user's background knowledge about the rates with which people perform a behavior; this prior can be constructed as a distribution over *different people*, each of whom do the behavior with a different rate. We designed our elicitation task to take advantage of the mixture-model representation of the prevalence prior used in Case Study 1. In particular, we assume, to a first approximation, that the distribution over prevalence can be represented as a mixture of those who tend to perform the action with a stable rate and those who do not perform the action. With the further assumption that, all else being equal, past is predictive of future behavior, we operationalize these two kinds of people as *people who have done the action before* and *people who have not the action before*. We design this experiment to measure participants' beliefs about the relative proportion of these two kinds of people (as a measure of the mixture parameter in the prevalence prior model) as well as the rate at which people (who have done the action before) do the action. We will assume for simplicity that people who have never done the action before will probably never do the action.[21]

Method.

Participants. We recruited 40 participants from Amazon's MTurk. Participants were restricted to those with U.S. IP addresses and who had at least a 95% work approval rating. The experiment took on average 12 min and participants were compensated $1.25 for their work.

Materials. To construct our stimulus set, we choose actions from five categories of typical human behaviors having to do with food and drug, work, clothing, entertainment, and hobbies. For each category, we created pairs or triplets of events that shared a superordinate action (e.g., *writing* poems vs. novels). The events were chosen to intuitively cover a range of likely frequencies. In total, 31 events were used. For a full list of the stimuli used in Experiments 2a–c, see Appendix D.

Procedure. For each event, participants were asked two questions, with different dependent measures. These questions were designed to measure the two components of the prevalence prior distribution. We anticipated there to be different beliefs about the rates and relative proportions of men versus women, so we asked about both genders separately. The two questions were schematically:

1. "How many {men, women} have *done action* before?"

Participants responded "N out of every J." by entering a number for N and choosing J from a drop-down menu (options: {1000–10 million}, incremented by 10×; default setting: 1000).

2. "For a typical {man, woman} who has *done action* before, how frequently does he or she *do action*?"

Participants responded "M times in K." by entering a number for M and choosing K from a drop-down menu (options: {week, month, year, 5 years}; default setting: year).

For example, one set of prompts read: "How many women have smoked cigarettes before?"; "For a typical woman who has smoked cigarettes before, how frequently does she smoke cigarettes?" Participants answered both questions for both genders on each slide (four questions total per slide, order of male/female randomized between-subjects), and every participant completed all 31 items in a randomized order. The difference in meaning of these questions was explained to participants on an instructions page before the experimental trials and tested for recall on a subsequent trial. Participants responded to this attention check by selecting an option from a drop-down menu consisting of four options (one correct description of the questions and three distractors).

Data analysis and results. All participants responded correctly to both questions in the attention check trial, so all collected data were used in the analysis. Question 1 elicits the proportion of people who have done an action before. We rescale this to be a number between 0 and 1, and model it as generated from a Beta distribution: $d_1 \sim \text{Beta}(\gamma, \xi)$. Question 2 elicits the rate with which a person (who has done the action before) does the action. We model this as generated by a log-normal distribution: $\ln d_2 \sim \text{Gaussian}(\mu, \sigma)$. Each item was modeled independently for each gender. We learned about the credible values of the parameters by running MCMC for 100,000 iterations, discarding the first 50,000 for burn-in.

The priors elicited cover a range of possible parameter values as intended (Figure 7A). We observe a correlation in our items between the mean % of Americans who have done action before (Question 1) and the mean log-frequency of action (Question 2) ($r_{1,2}(62) = 0.73$). Items in our data set that tend to be more popular actions also tend to be more frequent actions (e.g., *wears socks*) and vice versa (e.g., *steals cars*), though there are notable exceptions (e.g., *plays the banjo* is not popular but done

[21] It is likely that more than just these two possibilities are represented in people's intuitive theories, corresponding to individuals with additional traits or demographics, and influenced by the types of people a speaker knows and interacts with. We assume here a simple two-component structure so as to not make the specification of the prior overly complex.

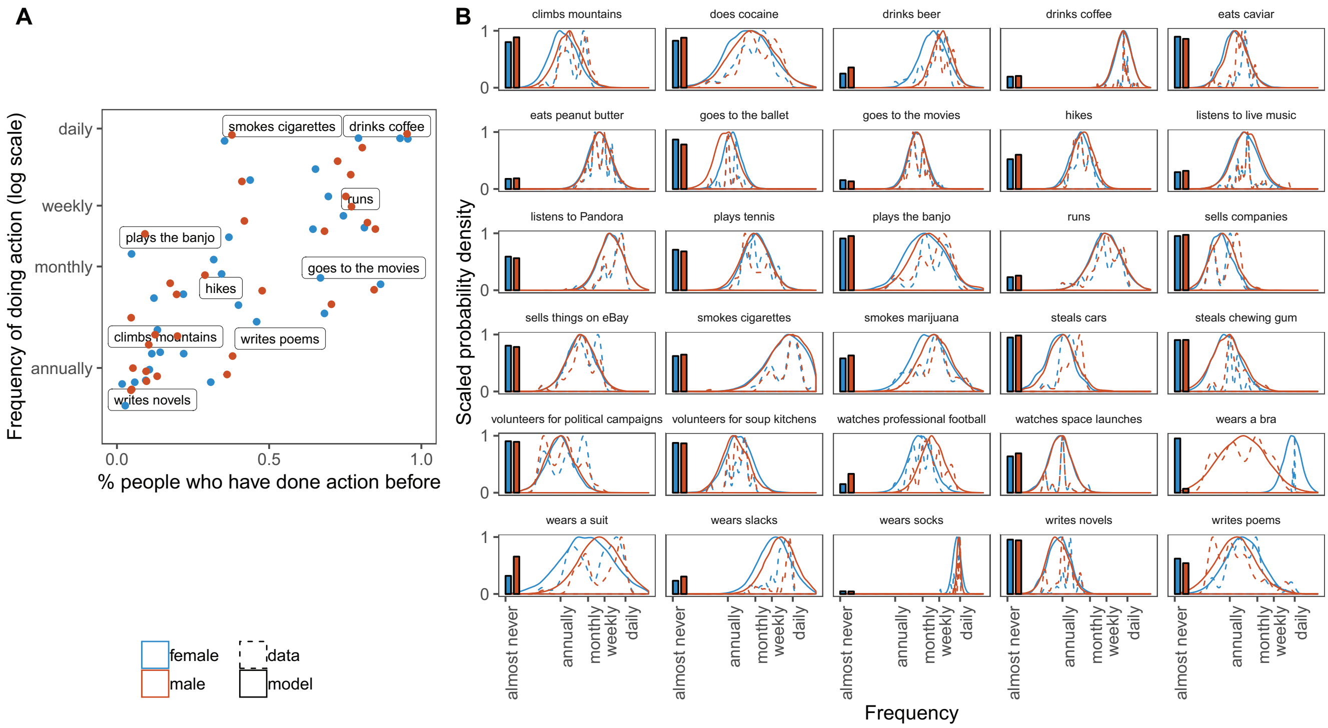

Figure 7. Prevalence priors for events (Experiment 2a). (A) Maximum A-Posteriori (MAP) estimates of parameters of prevalence priors for the 31 items in Experiment 2. Items cover much of the range of possible parameter values. (B) Reconstructed prevalence priors. To show frequencies for events that are very rare across people (e.g., *writes novels*), extremely low frequencies (*almost never*) are omitted. Instead, height of the bars on left denote the MAP values of the mixture component (in terms of *proportion of people who have never done action before*), reflecting the (inverse) popularity of the event across people. See the online article for the color version of this figure.

frequently when done at all, as is *smokes cigarettes*; *goes to the movies* is a popular activity though not done particularly often). This diversity is relevant because the speaker model (Eq. 3) will endorse habitual sentences (e.g., "Sam goes to the movies" vs. "Sam goes to the ballet") contingent on these details of the prior distribution.

To generate prevalence prior distributions, we built a Bayesian mixture-model for this prior elicitation task, analogous to that used in Case Study 1 (Experiment 1b). The only difference is that we estimate the mixture component ϕ directly from responses to Question 1. We assume that those who have not done the action before will probably not do the action in the future. With these assumptions, the prevalence distribution is given by:

$$\phi \sim \text{Beta}(\gamma_{Q1}, \xi_{Q1})$$
$$\ln p \sim \begin{cases} \text{Gaussian}(\mu_{Q2}, \sigma_{Q2}) & \text{if Bernoulli}(\phi) = \text{T} \\ \text{Delta}(p = 0.01) & \text{if Bernoulli}(\phi) = \text{F} \end{cases} \quad (4)$$

Figure 7B shows example reconstructed priors. In addition to specifying the correct way to combine our two prior-elicitation questions, using this inferred prior resolves two technical difficulties. First, it smooths effects that are clearly results of the response format. For example, a very common rating for certain events is *1 time per year*. Presumably participants would be just as happy reporting *approximately* 1 time per year (e.g., on average, 1.2 times per year); the raw data does not reflect this because of demands of the dependent measure. Second, this methodology better captures the tails of the prior distribution (i.e., very frequent or very infrequent rates) that have relatively little data and need to be regularized by the analysis. Now that we have modeled the prevalence prior data, we see whether our endorsement model can accurately predict endorsement rates for habitual sentences about these actions.

Experiment 2b: Habitual Endorsements

In this experiment, we elicit human endorsements for generalizations about events (*habituals*; e.g., "Mary smokes cigarettes") while manipulating the frequency with which the referent event occurs (e.g., how often Mary smokes cigarettes).

Method.

Participants. We recruited 150 participants from MTurk. To arrive at this number, we performed a Bayesian precision analysis to determine the minimum sample size necessary to reliably ensure 95% posterior credible intervals no larger than 0.3 for a parameter whose true value is 0.5 and for which the data are a two-alternative forced choice. This analysis revealed a minimum sample size of 50 per item; because participants only completed about one third of the items, we recruited 150 participants. The experiment took 4 min on average and participants were compensated $0.55 for their work.

Materials. Each event from Experiment 2a was paired with between two to four frequencies, for which the habitual statement would be evaluated. Frequencies were presented in terms of a character performing the action "three times in the past *time interval*." We chose to always have the character perform the action three times to provide a strong test of a baseline hypothesis that the habitual encodes the person has done the action several (at least three) times in the past.

Different time intervals were chosen for each event to maximize the variability of responses within each item. Specifically, we used the endorsement model to generate predictions based on the prior elicitation data (Experiment 2a) for each item, and choose between two and four time intervals across which maximal variability was predicted. For example, relatively high frequencies were chosen (e.g., time intervals of weeks and months) for items expected to occur rather often (e.g., *runs*); for an item that was expected to occur infrequently (e.g., *climbs mountains*), lower frequencies were chosen (e.g., time intervals of years or longer) because the model predicted that much of the variability in endorsement would occur in those respective ranges. In total, 93 unique items were created by pairing frequencies with events. The full list of items and frequencies can be found in Appendix D.

Procedure. On each trial, participants were presented with a *past frequency statement* for a given event of the form: "In the past {week, 2 weeks, month, 5 years, . . .}, Person did *X* 3 times." For example, "In the past month, Bill smoked cigarettes 3 times." Participants were asked whether they agreed or disagreed with the corresponding habitual sentence: "Person does *X*" (e.g., "Bill smokes cigarettes."). Participants completed 37 trials, which were composed of the 31 items from the prior elicitation task randomly paired with either a male or female character name. Six of these items were then also paired with a name of the opposite gender (e.g., participants rated both a female character and a male character who drank beer). These were used for an exploratory analysis on differences in endorsements by gender of the target character.

Results. The goal of this experiment was to elicit variability in habitual endorsements. Consistent with this goal, we found habitual sentences were endorsed for a wide range of frequencies. When actions are very infrequent (three times in a 5-year interval), habituals can receive strong agreement (e.g., *writes novels*, *climbs mountains*). When actions are relatively frequent (e.g., three times in a 1 month interval), habitual sentences can receive less than full endorsement (e.g., *wears socks*, *drinks coffee*). In our data, actions completed with a relatively high frequency (three times in a one week interval) receive at a minimum 75% endorsement, though there is still variability among them (e.g., between 10 and 25% disagree that people who wore a watch or wore a bra three times in the past week *wear a watch* or *wear a bra* habitually). Finally, we observe that none of our items receive less than 25% endorsement (i.e., a maximum of about 75% of participants disagree with the habitual utterances), reflecting the fact that these statements are not altogether false even though the action may be done very rarely.

Endorsement model comparison. In an exploratory analysis, we found no differences between endorsements of the habitual of characters with male and female names, and overall, the mean endorsements by gender were strongly correlated $r(93) = 0.91$. Endorsements are even more highly correlated for the six events we anticipated differences by referent-gender: $r(21) = 0.97$. This lack of a difference may be because the felicity of habitual sentences depends on a comparison with individuals of both genders (i.e., habituals are evaluated with respect to *other people*, not just other men or other women). Less interestingly, the lack of a difference may be the result of gender being not very salient in our paradigm, perhaps because the names used were not sufficiently gendered.

We now turn to our model-based analyses to better understand the endorsement data and the contribution of our model. For all analyses, we collapse across gender of the referent character for endorsement judgments. Parallel to our analysis of generic language endorsements, we articulate a set of simple regression models and a fixed-threshold alternative to our uncertain-threshold endorsement model. Analyses that use the prevalence prior distribution $P(p)$ (all models except "referent frequency" regression) use a 50% mixture of the inferred priors for each gender to construct a single prevalence prior distribution. Parallel to our analysis in Case Study 1, we model uncertainty in the input measurements by bootstrapping the data for the regression models and constructing joint-inference, Bayesian data analytic models for the information-theoretic endorsement models (fixed-threshold and uncertain-threshold; see Supplementary Model Criticism in Appendix C for a justification.).

Referent frequency. To understand the role of frequency in habitual endorsement, we use the frequency supplied to the participants in our experiment as a predictor in a linear model. This model predicts the same endorsement level for two actions done with the same frequency. Obvious counterexamples exist in our data set: While participants are willing to endorse that a person ". . . climbs mountains" having done it three times in the past year, they are less willing to say that a person ". . . hikes" and not willing to say that a person ". . . runs." Some actions done with a relative high frequency (e.g., three times in a month) do not receive full endorsement (e.g., *smokes cigarettes*; Figure 8A). Overall, the referent-frequency (in log-scale) predicts only a fraction of the variability in responses ($r^2(93) = 0.325$; $MSE = 0.0355$). In addition, for actions that are done on the time scale of years or longer (lower median of frequency), referent frequency no longer explains endorsements ($r^2(50) = 0.0662$; $MSE = 0.0477$). The prevalence baseline does appreciably worse in this data set in comparison with the generics data set (Case Study 1) because we were able to independently manipulate the referent-frequency separate from the prevalence priors, which we could not do for generics about familiar categories.

Distinctiveness and referent frequency. In our empirically elicited priors, items differ in the proportion of people who have done the action before (the mixture parameter of the mixture model; Figure 7A *x*-axis). This mixture parameter is a major contributor to the mean of the prevalence prior distribution and, thus, relates to the cue validity of a particular feature for a particular individual (see Appendix A). Thus, we take this parameter as an index of the distinctiveness of the action, analogous to cue validity in the case of generic language. We construct a regression model that treats endorsement as a linear combination of the frequency given to participants and participants' responses to the question about the mixture parameter ϕ (i.e., the proportion of people who have done the action before) as an index of distinctiveness.

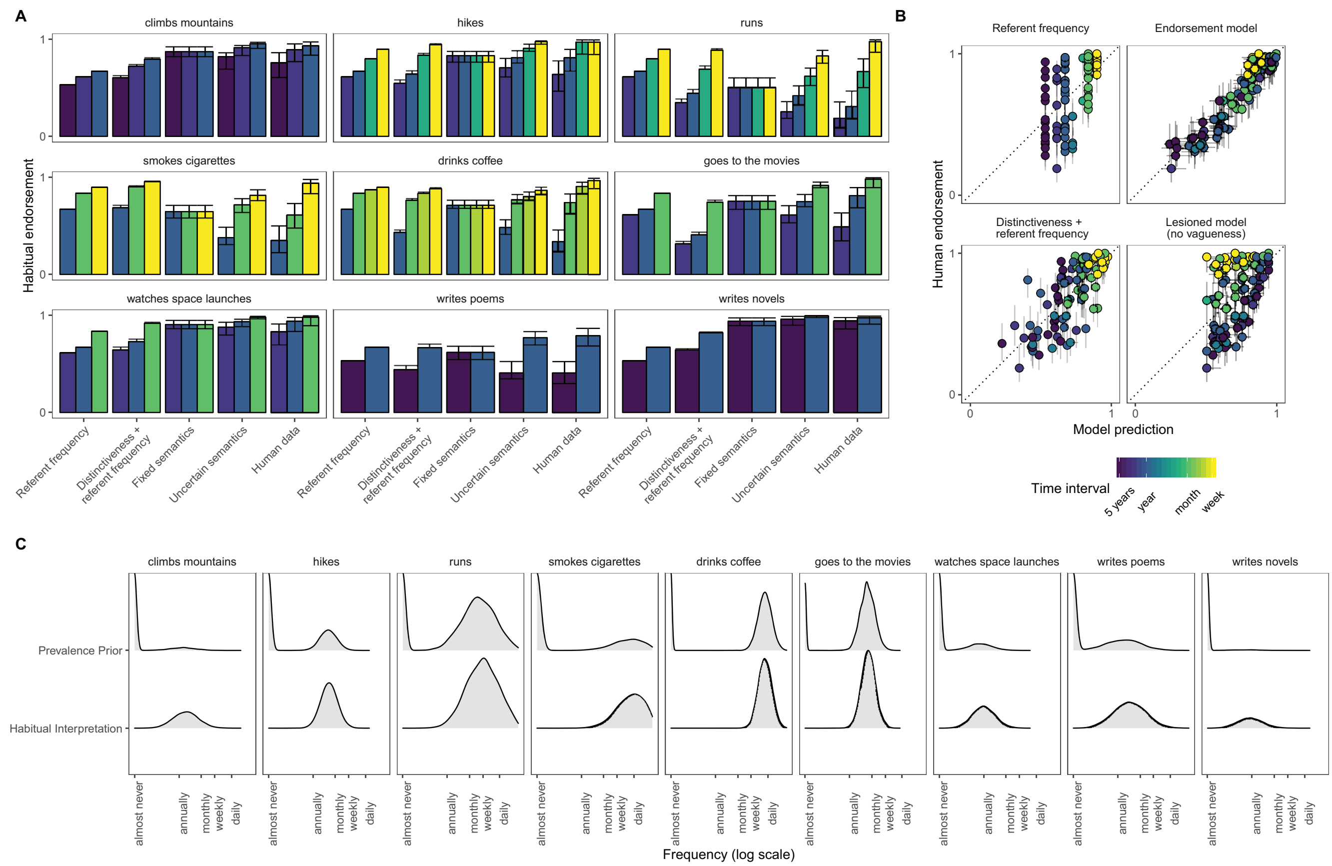

Figure 8. Endorsing generalizations about events. (A) Endorsements by human participants and four models for nine events given different frequencies of action. (B) Model fits for all 93 habitual sentences by each model. (C) Nine example frequency priors and posteriors upon hearing the habitual. These distributions are inferred using both data sources from Experiments 2a and 2b. See the online article for the color version of this figure.

This model is able to explain more of the variance in endorsements ($r^2(93) = 0.583$; $MSE = 0.022$). It can differentiate events done with the same frequency (e.g., *writing poems* vs. *novels*, three times in the past 5 years) by increasing endorsement of the more rare action (*novels*). Still, this model fails to capture fine-grained differences in endorsement. For example, going to the movies is a relatively nondistinctive action (many people do it) and going three times in a year is not very frequent and, yet, people still strongly endorse the habitual (mean endorsement and 95% CI 0.81 [0.69, 0.89]), while this regression model predicts quite lower judgments (0.41 [0.41, 0.43]). On the other hand, playing the banjo three times in the past 2 years is not strong evidence for the habitual, according to participants (0.45 [0.33, 0.59]). Nevertheless, because playing the banjo is a distinctive action, the regression model wants to endorse the habitual strongly in this case (0.75 [0.75, 0.76]).

Lesioned model (no vagueness). We next examine an information-theoretic endorsement model based on a fixed-threshold semantics. This model is identical to the full endorsement model, but is lesioned to not have vagueness or uncertainty about the meaning. A fixed-threshold model commits the habitual to conveying literally that a person *does the action with some frequency*, and that threshold on frequency is the same for all actions. As in Case Study 1, we incorporate this model into a Bayesian joint-inference model to infer the fixed-threshold and simultaneously predict both the prior data and the endorsement data (for more details on model implementation, see Appendix C). We assume the referent-prevalence p being conveyed by the endorsement model (Eq. 3) is the frequency provided to participants (e.g., three times in the past year). Additionally, to account for statements that would be literally false under this model (frequencies that fall below the fixed threshold), we include an additional noise parameter, as we did for the fixed-threshold model in Case Study 1. To learn about the credible values of the model's parameters and generate predictions given those inferred parameter values, we collected 2 MCMC chains of 100,000 iterations, discarding the first 50,000 iterations for burn in.

The data analytic model infers that a low threshold is likely: the Maximum A-Posteriori threshold and 95% CI in units of number of times per year is 0.01 [0.01, 0.37]. Compare this with the lowest referent-frequency used in our data set: 0.6 times per year (three times every 5 years). Thus, all of the utterances evaluated under this fixed-threshold model were literally true. As a result, the lowest endorsement this model can apply to an utterance is 0.5, because both the habitual and silence are always true. The fixed-threshold model exhibits a small dynamic range of endorsements, similar to the referent-frequency model (Figure 8B).

The fixed-threshold habitual updates the interpreter model's prior beliefs differentially depending on the item. For instance, because of the distinctiveness of *climbs mountains*, the fixed-threshold endorsement model fully endorses the habitual ("Mary climbs mountains") even at a low frequency (Figure 8A). However, the model cannot differentiate among different frequencies of doing the same action, because they are all above the truth-functional threshold. It is equally true that a person does an action with nonzero frequency for any frequency greater than zero, analogous to how "Some dogs are friendly" is equally true whether 20, 50, or 80% of dogs are friendly.

Overall, the fixed-threshold model is able to predict only a fraction of the variance in human endorsements ($r^2(93) = 0.299$; $MSE = 0.0369$). The model does this by inferring that 10% [1, 19] of the data are noise and that the speaker optimality parameter is 0.77 [0.70, 1.25].

Uncertain threshold model. We used the same data analytic approach for the uncertain threshold endorsement model and performed the same Bayesian statistical inference over the model to learn about its parameters and predictions. Again, this model has two fewer parameters than the fixed-threshold model (no fixed-threshold parameter and no extrinsic noise process). As shown in Figure 8B, the uncertain-threshold endorsement model does a good job of accounting for the variability in responses ($r^2(93) = 0.894$; $MSE = 0.00598$), including actions done on the time scale of years or more ($r^2(50) = 0.903$; $MSE = 0.00617$).

Figure 8C provides insight into how the uncertain-threshold model is able to match human judgments. The endorsement model simulates how an interpreter would understand the habitual sentence. A habitual is interpreted relative the prior distribution over frequencies, and the comparison between the frequency implied by the habitual versus staying silent results in different frequencies at which the generalization is good to assert. Climbing mountains three times in the past year is good evidence that you *climb mountains* because it is approximately the frequency that a listener would infer given the utterance; going for a hike three times in the past year is correspondingly less convincing that you *hike*; and if you went for a run three times in the past year, you not a person who *runs*. Only the uncertain-threshold model is able to draw these subtle distinctions.

Discussion. Habitual language exhibits context-sensitivity directly parallel to that of generic language (Case Study 1). Habituals are endorsed for a wide range of frequencies, but show systematic patterns relative to the prior distribution of frequencies, as formalized by the uncertain-threshold model. Again, we articulated a number of alternative models and found that only the underspecified threshold model was able to explain the variability in endorsements.

In this case study, we manipulated rather than measured the referent frequency (e.g., the frequency with which a person drinks coffee). By manipulating the target frequency, we have shown that it is causally related to habitual endorsements in the way predicted by our model and in a way that a fixed-threshold model cannot account for. The relationship is not linear, however; habitual endorsements vary in complex ways that reflect interpreters' prior knowledge about the event in question.

In Experiment 2b, participants were given a statement about how often a person has done the action in the past and asked to judge the corresponding habitual statement. This design potentially confounds an important distinction for the language of generalization: Does the prevalence communicated by a generalization indicate an objective, past frequency or a subjective, future expectation? In Experiment 2c, we investigate this question by teasing apart past from predictive frequency and measuring its influence on habitual endorsement.

Experiment 2c: What Is Prevalence?

While past frequency is often a good indicator of future tendency, the future is under no obligation to mimic the past. Does habitual language communicate probabilities in terms of past frequency or future expectations? On one hand, speakers can only be certain about what has happened in the past. On the other hand, it is important for speakers to be able to convey their predictions of what they believe will be the case in the future.

People can change their behavior abruptly because of a variety outside events (e.g., developing an allergy) or intend to do an action without actually completing it (e.g., by making a resolution). We introduce these causal events into our experimental paradigm to measure their influence on endorsement. To provide the appropriate model-based analysis, participants in one condition make a prediction about the future (*predictive frequency*). In another condition, participants decide whether or not to endorse the habitual sentence (as in Experiment 2b). We then compare two uncertain-threshold models: one which uses participants' ratings of predictive frequency as the referent prevalence and one which uses the past frequency (as was done for Experiment 2b). In addition, we compare to a baseline linear model that uses only the predictive frequency (no priors) to model endorsement.

Method.

Participants. We recruited 270 participants from MTurk, using the same criteria as Experiment 2b. There were 120 assigned to the *predictive frequency* condition and 150 were assigned to the *habitual endorsement* condition. The experiment took on average 3.50 min (*predictive frequency*) and 2 min (*habitual endorsement*). Participants were compensated $0.40.

Materials. The events used were a subset of those used in Experiments 2a and 2b (21 of the original 31). In addition, we crafted statements that were intended to either increase the frequency (*enabling*; e.g., "Yesterday, Bill bought a pack of cigarettes.") or decrease the frequency (*preventative*; "Yesterday, Bill quit smoking.") of the event in the future. To increase the potential variability of responses across the experimental conditions, participants only saw the frequencies that led to the most intermediate endorsement of the habitual in Experiment 2b. We did not include separate trials for both male and female names for the select items we did in Experiment 2b, because we saw no differences in their endorsements of the habitual. See Appendix D for a full list of the items and frequencies used, as well as the enabling and preventative information.

Procedure. The procedure was identical to Experiment 2b except for the inclusion of a second sentence on a subset of trials (*preventative* and *enabling* trials). On all trials, participants were presented with a past frequency sentence (same as Experiment 2b). Additionally, trials either included preventative information, enabling information, or no additional information (identical to Experiment 2b), in equal proportions. See Table 2 for example trials.

In the predictive frequency condition, participants were asked "In the next *time interval*, how many times do you think *person does action*?", where the *time interval* was the same as given in the past frequency statement. In the habitual endorsement condition, participants were asked if they agreed or disagreed with the corresponding habitual sentence (as in Experiment 2b).

Predictive frequency results. Figure 9B shows the mean predicted future frequency as a function of the past frequency given to the participant and the type of causal information given. We observe in the baseline condition that future frequency perfectly tracks past frequency ($r(21) = 0.994$). That is, participants believe if a person smoked cigarettes three times last month, they will smoke cigarettes three times next month. This result implies that our model makes identical predictions for Experiment 2b whether the referent is past frequency or expected future frequency (indicating, as expected, that we must look to new data to distinguish these models). Critically, we observe the preventative information strongly decreases and the enabling information slightly increases predicted frequency (Figure 9B, white/lightest and dark green/darkest dots).

We confirmed these observations using a linear mixed-effects model, predicting the log-transformed responses from the log-transformed past frequency and the experimental condition (baseline, preventative, and enabling). To account for participant and item variability in this analysis, we also included random effects of intercept and condition for both participants and items. Confirming that our manipulation worked as intended, the preventative information led to significantly lower predictions for future frequency, relative to the baseline condition ($\beta = -3.18$; $SE = 0.27$; $t = -11.96$). There was also a tendency for the enabling information to lead to higher predictions for future frequency, relative to baseline ($\beta = 0.96$; $SE = 0.11$; $t = 8.54$). Finally, past frequency was a significant predictor of predicted future frequency ($\beta = 1.01$; $SE = 0.02$; $t = 40.80$).

Habitual endorsement results. There is a clear and consistent negative effect of preventative information on endorsements for the habitual sentences (Figure 9A; white bars). When collapsing across items, the Bayesian Maximum A-Posteriori estimate and 95% highest probability density interval for the true endorsement probabilities per condition are: baseline = 0.85 [0.83, 0.87], enabling = 0.90 [0.88, 0.92], preventative = 0.29 [0.26, 0.32]. Still, frequency—even predictive frequency—does not perfectly explain the endorsements ($r^2(63) = 0.518$; $MSE = 4.29$; Figure 9B).

We use our formal model to test whether past or predictive frequency matters for endorsement. To formalize the predictive frequency speaker model, we use the mean predictive frequency as the referent-prevalence p that the endorsement model (Eq. 3) aims to convey. The past frequency model is constructed using the past frequency supplied to participants as the referent-prevalence. We analyze this model in the same Bayesian data analysis regime as for our previous models. We use the same priors over the parameters as before and learn about the posterior distribution by collecting three independent MCMC chains of 100,000 iterations (removing the first 50,000 for burn-in). Figure 9C shows the resulting model predictions for the past frequency and the predictive frequency endorsement models. Participants' judgments of the habitual statements was indeed influenced by the causal manipulations in the way predicted by the endorsement model that uses the predictive frequency as the referent prevalence ($r^2(63) = 0.931$; $MSE = 0.00594$). The model based on past frequency does not make different predictions for the different causal manipulation conditions and does a poor job at explaining the endorsements ($r^2(63) = 0.0334$; $MSE = 0.0833$). This result strongly suggests that prevalence represents a predictive belief about the future.

Discussion. Habitual language conveys generalizations about events. Our model decides if a habitual sentence is a pragmatically useful way to describe the rate at which a person does an action, taking into account a naive interpreter's prior beliefs about the event (measured in Experiment 2a). Our computational model endorses statements that communicate generalizations about events with the same sensitivity to context and frequency that people exhibit (Experiments 2b and 2c). In Experiment 2b, we varied the type of event and the past frequency with which the person did the action, and found graded endorsements of the corresponding habitual sentences. By manipulating (rather than measuring) the referent frequency, we showed how alternative models were unable to account for the gradience in endorsement. In particular, we show that prior knowledge in an information-theoretic, communicative model is not sufficient to produce gradience in endorsement: The fixed-threshold model, which has these components, does not make different predictions for different frequencies. Only our uncertain threshold model was able to precisely account for the wide range of endorsements.

In Experiment 2c, we further investigated the nature of the underlying prevalence scale by introducing causal information that

Table 2
Example Stimuli Used in Experiment 2c

Habitual	Baseline	Preventative	Enabling
John smokes cigarettes.	In the past month, John smoked cigarettes 3 times.	In the past month, John smoked cigarettes 3 times. Yesterday, John quit smoking cigarettes.	In the past month, John smoked cigarettes 3 times. Yesterday, John wanted a smoke and bought a pack of cigarettes.
Tina volunteers at soup kitchens.	In the past 5 years, Tina volunteered for soup kitchens 3 times.	In the past 5 years, Tina volunteered for soup kitchens 3 times. Yesterday, Tina grew disillusioned with the soup kitchen system and wants nothing to do with it anymore.	In the past 5 years, Tina volunteered for soup kitchens 3 times. Yesterday, Tina researched a new soup kitchen in the area and is going to volunteer with them.

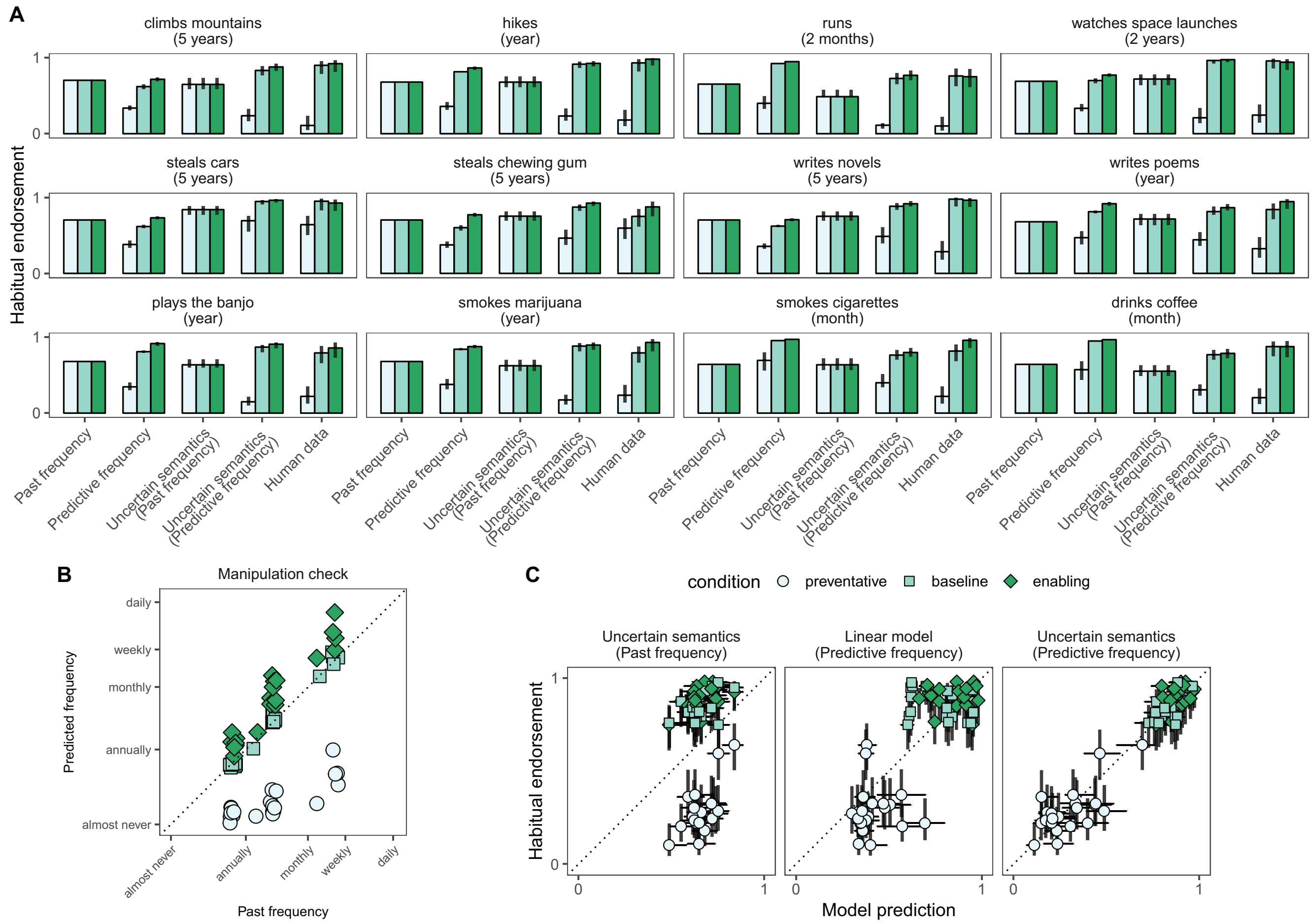

Figure 9. (A) Example empirical and predicted endorsements for habituals in three conditions. (B) Predicted frequency as a function of past frequency and condition manipulation (enabling, preventative, and baseline). (C) Model fits for the uncertain threshold model using past frequency, linear model based on future frequency, and uncertain threshold model using future frequency. See the online article for the color version of this figure.

enabled and prevented future occurrences of the action. We used the empirically measured predicted future frequency as the object of communication for our endorsement model. We found that the endorsement model that seeks to communicate its predictions (rather than its observations) is a better model of habitual endorsements under these situations. That is, habitual language (and generalization language more generally) is fundamentally about conveying people's predictive beliefs, not what has actually happened.

In these experiments, we introduced participants to novel actors and, by doing so, were able to directly manipulate the referent frequency. The kinds of events we used were familiar to participants (e.g., *running*) and, thus, we measured the prevalence priors for those events. In our final case study, we experimentally manipulate the prevalence priors, testing their causal influence over endorsements. In addition, we further we extend our theory to the language of causal relationships.

Case Study 3: Causal Language

Language about causal relationships manifests in generalizations. The utterance "Fire causes smoke" relates to "This fire caused this smoke" in a way analogous to how "John runs" relates to "John ran yesterday." We explore this hypothesis in our third case study: causal language or *causals* (e.g., "A causes B").

The problem of *causal induction*—knowing that one thing causes another—has been studied extensively in human psychology (Cheng, 1997; Cheng & Npvick, 1992; Griffiths & Tenenbaum, 2005, 2009). Classically, this is cast as a problem of inducing an unobservable relation (*type causation*) from observable events or contingency data. We take a different approach, examining *type causation* by the language used to describe it (e.g., "A causes B"). We explore the idea that such language conveys a generalization about *token* or *actual causation* (e.g., "A caused B, in this instance") and that our theory of the language of generalization extends in a natural way to describe *causal language*. Ascribing causation to an individual event (token or actual causation) is itself a complex, inferential process (e.g., depending on counterfactual reasoning; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015), which we do not try to model here.

In this article, we posit that prevalence priors are a mediating representation between abstract conceptual structure and the language of generalization. In this last set of experiments, we explicitly test the relationship between the prevalence priors and endorsements of generalizations about causes by manipulating the priors.

Experiment 3a: Manipulating Prevalence Priors

In this experiment, we manipulate participants' background knowledge, measuring these beliefs to check whether the manipulation was successful. Experiment 3b (*causal endorsement*) will then use a very similar experimental procedure in exploring the language of causal generalizations.

Method.

Participants. We recruited 160 participants from Amazon's MTurk. Participants were restricted to those with U.S. IP addresses and who had at least a 95% work approval rating. The experiment took on average 1.70 min and participants were compensated $0.50 for their work.

Materials. Participants were told a story of a scientific experiment testing different substances to produce an effect (either to make animals sleepy or make plants grow tall). Our cover stories were constructed so that the potential cause could have some plausible intuitive mechanism that could give rise to the property (e.g., a naturally occurring herb causing animals to be sleepy). The two cover stories can be seen in Table 3.

Participants were then shown *previous experimental results* which followed one of four distributions represented as a table of numbers. In two of the conditions, participants saw results that came from a single underlying distribution (*common* conditions). In one of these conditions, all causes produced a strong effect (average efficacy approximately 98%; the *common strong* condition). In the second of these conditions, all causes produced a weak effect (average efficacy approximately 20%; the *common weak* condition). The two other conditions used distributions in which some experiments resulted in either no or very few successes (i.e., produced 0, 1, or 2 successes), and others that either had strong or weak effects as above. These are the *rare strong* and *rare weak* distributions.

Procedure. The experiment was a single trial: Each participant saw only one cover story with one distribution of previous experiments. Participants were told that they were an astronaut-scientist on a distant planet trying to figure out how some system works (i.e., how to make a certain kind of animal sleepy with different herbs or how to make a plant grow tall with different fertilizers). The story for the *sleepy animals* condition read:

> You are an astronaut-scientist exploring a distant planet. On this planet, there are animals called cheebas and your team of scientists wants to figure out how to make these animals sleepy. Your team runs experiments trying to make cheebas sleepy with different naturally occurring herbs. The results are shown below:

Participants then clicked a button to show the results of the experiments, which appeared one at a time in a random order (following a particular distribution). Experimental results were also described linguistically (e.g., "Your team gave herb A to 100 different cheebas. Of those 100 treated, 98 cheebas were made sleepy.") as well as displayed in a table showing the number of successes (e.g., "animals made sleepy") per number of attempts (always 100 per experiment; Figure 10A). We described the results of these individual experiments using token-level causal language (e.g., "98 cheebas were made sleepy") to imply that actual causation occurred in these cases. Participants see the results of 11 experiments, though they are told the results of one experiment were lost and a "?" was placed in the table (Figure 10B). (These lost results would be found in the *causal endorsement* task, Experiment 3b.) After participants viewed the results of the 10 experiments (and one missing experiment), they are told to review the results of the experiments before continuing (Figure 10C).

Upon clicking the continue button, the table of experiment results is removed and participants are told that five more experiments were conducted that day and asked to predict the results of those experiments (Figure 10D). Participants were given five slider bars ranging from 0–100 to rate the number of predicted successes out of 100 attempts. After responding, participants then completed an attention check survey where they were asked what the team of scientists was investigating (choosing a response from a drop-down menu with 12 options) and to input one of the numerical results they saw on the previous screen. This attention check served to confirm that participants had encoded both relevant aspects of the experiment (the domain and the frequencies).

Results. Twenty participants were excluded from the analysis for failing to answer both of the attention check questions correctly, leaving a total of 140 responses for analysis. The empirically elicited distributions of responses were not appreciably different for our two cover stories (herbs making animals sleepy, fertilizer making plants grow tall) and, thus, we collapse the data across these two stories. The distributions that resulted from participants predicting the causal efficacy of the new substances are shown in Figure 11A. As is visually apparent, the empirical prevalence distributions differ between conditions and nicely recapitulate the distributions supplied in the different experimental conditions, suggesting that the manipulation does indeed change participants' representations of what probabilities are likely to occur in each experimental condition. This diversity is important

Table 3
Cover Stories and Example Evidence Statements for the Two Sets of Materials Used in Experiment 3

	Plants	Animals
Cover story	On this planet, there is a plant called feps and your team wants to figure out how to make these plants grow tall. Your team runs experiments trying to make feps grow tall with different fertilizers.	On this planet, there are animals called cheebas and your team of scientists wants to figure out how to make these animals sleepy. Your team runs experiments trying to make cheebas sleepy with different naturally occurring herbs.
Evidence statement	Your team gave fertilizer B to 100 different feps. Of those 100 treated, 2 feps grew tall.	Your team gave herb C to 100 different cheebas. Of those 100 treated, 98 cheebas were made sleepy.

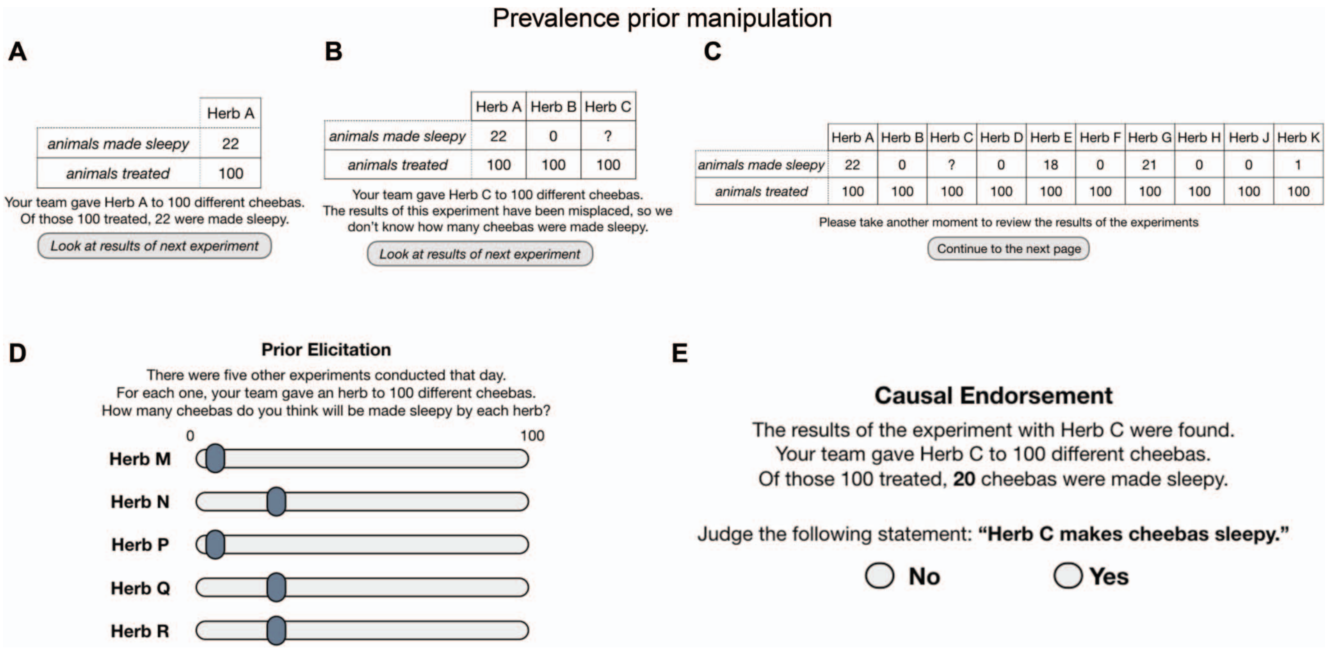

Figure 10. Overview of Experiment 3. (A–C) Results of previous experiments are shown one at a time, described in text and displayed in a table. One of the results was lost. Participants are asked to review previous results once all displayed. (D) Prior elicitation task: Participants predict the results of the next five experiments. (E) Causal endorsement task: Results of previously lost experiment are found and participants are asked to evaluate the causal generalization. See the online article for the color version of this figure.

because the model of generalizations predicts differences in endorsement—for the same referent prevalence—depending on these priors.

Experiment 3b: Causal Endorsements

In this experiment, we tested whether the manipulated priors of Experiment 3a are causally related to the endorsement of causal statements. Most of the experimental design was identical to that of Experiment 3a.

Method.

Participants. We recruited 400 participants from Amazon's MTurk. Participants were restricted to those with U.S. IP addresses and who had at least a 95% work approval rating. None of the participants had participated in Experiment 3a. The experiment consisted of one trial and took on average 1.40 min; participants were compensated $0.25 for their work.

Procedure and materials. The materials were the same as in Experiment 3a. The first part of the experimental trial was the same as in Experiment 3a (the table of previous experiments; Figure 10A–C). Upon continuing beyond the first part of the trial, the table of results and background story were removed from the screen and the participant is told that the results of the "lost experiment" were found (the experiment with a "?" in the table).[22] The results are reported to the participant in terms of how many out of 100 of the attempts were successful. Participants saw one of two reported frequencies: 20 or 70% (randomized between-subjects). Participants were then asked to judge the causal sentence (e.g., "Herb X makes the animals sleepy") by either clicking "Yes" or "No" (Figure 10E). This allows us to test whether endorsement of a causal sentence for a given actual frequency is affected by the causal priors induced by our manipulation. After responding, participants completed the same attention check as Experiment 3a.

Results. Forty-two participants were excluded from the analysis for failing to answer both of the attention check questions correctly, leaving a total of 358 responses for analysis. As in our other analyses of endorsement responses, we computed the Bayesian Maximum A-Posteriori (MAP) estimate and 95% highest probability density interval of the true population probability of endorsing the statement, assuming a uniform prior. These are shown for the different experimentally manipulated priors and referent prevalences in Figure 11B.

As predicted by our model, endorsements for a causal statement were sensitive to the referent prevalence of causal events and, critically, to the background distribution of other causes. When many other causes produced the effect very reliably (*common strong* condition), very few participants endorsed the causal statement for a causal frequency of 0.2, and were at chance when the causal frequency was 0.7 (Figure 11, blue bars). By contrast, when many other causes failed to produce the effect and those that did were not very reliable (*rare weak* condition; green in figure), at least half of participants endorsed the causal statement for a cause

[22] We chose to have the scientists report on an earlier "lost experiment" to suggest a binomial generative process for the experiments wherein the scientists planned to perform 11 experiments, as opposed to alternative design wherein an 11th experiment is reported. Such a continuation of the series of experiments could imply a generative model following a geometric distribution where the scientists repeat the experiment until they reach one that is successful.

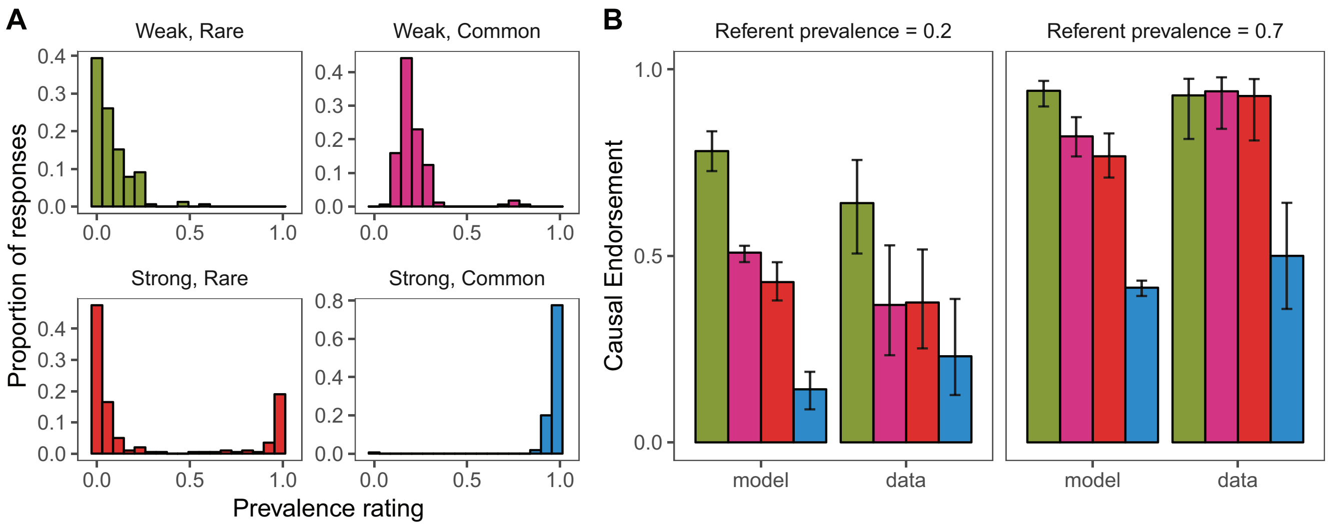

Figure 11. Endorsing generalizations about causes. (A) Empirical prevalence prior distributions elicited following prior manipulation cover story (Experiment 3a). (B) Endorsement model predictions and human elicited endorsements for four manipulated prevalence prior distributions (colors) and two referent prevalence levels (facets). See the online article for the color version of this figure.

with causal prevalence of 0.2, and were at ceiling when the prevalence was 0.7. The other two conditions (*rare deterministic* and *common weak*) led to endorsements intermediate between these two conditions. These effects were predicted by our model with strong quantitative accuracy ($r^2(8) = 0.835$; $MSE = 0.0123$).

Discussion. In our third case study, we applied our model to generalizations about causal events, without any changes. In this domain, we successfully manipulated participants' beliefs about the expected prevalence of a causal relationship in a domain (Experiment 3a). This was done using both unimodal (*common weak*, *common strong*) and bimodal (*rare weak*, *rare strong*) distributions. In Experiment 3b, we showed that these manipulated priors influenced endorsements of the corresponding causal statements. In addition to further demonstrating the generality of this theory, these experiments show that the prevalence prior $P(p)$ is causally related to endorsements of generalizations in language.

In these experiments, we used two cover stories that described plausible causal events: herbs making animals sleepy and fertilizers making plants grow tall. We chose these items because there was a plausible causal mechanism that could give rise to the property and these causal events could have ambiguous causal power associated with them (e.g., it is plausible that there are herbs that only weakly make animals sleepy and it is also plausible that there are herbs that almost deterministically make animals sleepy). These two features of the domains make them particularly amenable to manipulation. It is likely that other domain knowledge would interact with the experimentally supplied "experimental data" to form a hybrid belief distribution.[23] For example, physical causal systems (e.g., billiard balls hitting each other) could strongly induce near-deterministic notions of causality, analogous to our *strong* priors conditions. Causal systems that demonstrate surprising or *a priori* unlikely effects (e.g., liquids melting concrete) could induce rarity about the existence of a nonzero causal power, analogous to our *rare* prior conditions. Our theory would predict that differences in endorsement in these cases would be mediated by differences in the corresponding prevalence prior distributions.

General Discussion

The human species is remarkable not only because we can extract useful generalizations from observing the world, but because we can convey these generalizations to each other succinctly using language. Generalizations expressed in language (e.g., "John runs," "Dogs are friendly," and "Fire causes smoke") are a premier example of how simple statements—statements understood by even the youngest language users—can convey rich meanings and display complex sensitivities to context. We have argued that the core meaning of such linguistic expressions can be understood by three ingredients: Probability, vagueness, and context. That generics are vague does not preclude them from being treated by formal models. A vague threshold meaning operating over a speaker's inductive beliefs, formalized by predictive probability, leads to complex, quantitative interactions with background knowledge, which closely track human judgments.

We tested this theory by exploring its implications for the simplest judgments a person can make about a sentence: an endorsement or truth judgment. Truth judgments are the standard measurement for semantic theories, but are often limited to the intuitions of a single or a few trained theorists, often producing only binary or ternary judgments (i.e., true, false, and possibly indeterminate). Our model accounted for these standard linguistic intuitions about a number of philosophically puzzling generic

[23] If this were happening in our domains, we would expect this to show up in the results of Experiment 3a. Participants' predictions about the likely causal power of new causes would be expected to show a mixture of their abstract, intuitive theories and the experimentally supplied data.

statements (*Worked Examples*). We went beyond intuitions, however, measuring truth judgments of naive language users, which revealed substantial gradience in endorsements. These fine-grained, quantitative measurements pose real challenges for verbal theories of generics, which can only predict qualitative differences. The framework we propose is sufficiently precise to predict quantitative gradience, and we showed that the gradience in truth judgments depends in systematic ways on an interpreter's prior beliefs about prevalence. We provided strict tests of our formal model by both measuring and manipulating listeners' background knowledge as well as comparing our model to a number of alternatives. In each case, our model provided a strong quantitative fit to human judgments while other models fell short.

The fact that our model applies equally as well to generic and habitual language sheds light on a long-attested relationship. The analogy of habituality (e.g., "John runs") and genericity (e.g., "Dogs have four legs") has typically been assumed in the literature following the original suggestion by Carlson (1977). The relationship, however, has been never been empirically tested nor formally described. This article presents the first empirical evidence that the reasoning involved in understanding generic and habitual sentences are similar: The same uncertain threshold mechanism operating over prior beliefs explains judgments of both kinds of statements (Experiments 1 and 2).

In our model, an utterance conveying a generalization updates a listener's *a priori* beliefs about the prevalence of the property for that category. Because the prevalence prior includes the information necessary to compute cue validity (see Appendix A), the cue validity of the feature for the new category is added to the common ground between interlocutors. Thus, given the knowledge represented in a prevalence prior, generics communicate information both about the prevalence of the feature in the new category (via the literal semantics) as well as the cue validity of the feature for the new category. For example, if a listener hears "Alligators grow to be 10-feet long," they will update their personal beliefs both about the prevalence of 10-footedness among alligators as well as the probability that an animal is an alligator given that it is 10 feet long. To our knowledge, this is the first theory of generics that describes how the informational contribution of a generic relates to cue validity.[24] This also means that our computational framework makes predictions about the cue validity implied by different generalizations, without any further assumptions.

We propose that the language of generalization conveys predictive probabilities. Communicating probabilities might seem contrary to the long attested failures in human reasoning about probability (Tversky & Kahneman, 1974). Our theory suggests that the problems with understanding probabilities observed in classic, cognitive psychology paradigms are a problem in understanding explicit probabilities expressed using *numerical language*, a historically quite recent innovation (cf., Levinson, 1995). Rather than conveying probabilities explicitly, the language of generalization conveys them implicitly. In other words, we argue that the utterance "70% of birds fly" is a precise statement about how many birds fly in the way that "John is 6'3"" is a precise statement about the height of John, whereas "Birds fly" is a vague statement more akin to the utterance "John is tall." The latter statements are easier to process, understood at an earlier age, and may be more useful for human reasoning. Indeed, even infants are actually quite good with reasoning about probabilities, but in ways that are not explicit or tied to the numerical language of probability (Xu & Denison, 2009; Xu & Garcia, 2008).

In the rest of this discussion, we further discuss the contribution of our modeling framework to empirical and theoretical work on the language of generalization, elaborate further on the role of conceptual knowledge in our modeling framework, describe how our model relates to language understanding more generally including theories of vagueness, and sketch an argument for how our formal representation of the language of generalization could provide a relatively simple acquisition problem for a language learner.

Generics and Genericity

The statistical versus conceptual distinction for theories of generics, described in detail earlier in this article, is roughly coextensive with the distinction between theories of the truth conditions of generics and the mental phenomenon of *genericity*, respectively (Nickel, 2016). Truth conditions are unified criteria that all generics must satisfy to be true. Genericity, on the other hand, is thought to be a nonlinguistic, psychological phenomenon previously explained either in terms of pragmatics (Declerck, 1991) or metasemantics (Leslie, 2007; Liebesman, 2011; Nickel, 2016; Sterken, 2015). A full account of generic language must include both a theory of generics (i.e., truth conditions) and a complementary theory of genericity (Nickel, 2016).

Our model leverages the insights of formal semantics and computational cognitive modeling to articulate how an agent should update their beliefs based on a generic sentence. The Bayesian model provides a clean separation of the semantics of a generic statement (represented formally by a threshold function) from world knowledge (represented by the prevalence prior). We, thus, chart-out a single answer to the semantic question "what do generics mean?", while also formalizing how background knowledge influences generic understanding (a question about genericity). Previous theoretical accounts have either aimed to account for the context-sensitivity of generics by positing distinct semantic constructs (e.g., relative vs. absolute generics; Cohen, 1999) or positing a semantics that cannot be separated from world knowledge (Leslie, 2008). The modeling approach we take here demonstrates how the context-sensitivity of generics can emerge via the interplay of a stable (but vague) semantics and diverse background knowledge. Thus, this model is the first step towards unifying conceptual approaches to genericity with an inherently quantitative natural language semantics for generics.

Relation to Intuitive Theories

Our theory assumes that listeners and speakers use real-world knowledge of objects and events to use and interpret the language of generalization. It is this real-world knowledge (presently formalized in terms of a prevalence prior) where issues about domain specific beliefs, interpretation of properties (Prasada & Dillingham, 2006), overhypotheses (Leslie, 2008), and essentialism (Gelman, 2003) can all play a role. We hypothesize that all of this knowledge feeds into participants' predictions, measured through judgments about prevalence. That is, even though our underlying

[24] We are grateful to an anonymous reviewer for drawing our attention to this.

semantics is defined in terms prevalence of individual properties, this does not imply that the concepts involved in generic language understanding are defined merely by their properties. Instead, complex intuitive theories (that may reflect, e.g., conceptual role, see Goodman, Tenenbaum, & Gerstenberg, 2015) give rise to prevalence priors, which then yield judgments about generic statements. Thus, it is the output of these cognitive processes and conceptual understanding—intuitions about prevalence—to which this framework brings precision and clarity. Applying the same precision and clarity to the more abstract aspects of conceptual knowledge including property knowledge, essentialism, and so forth, is an important next step for this line of work. Probabilistic causal models (Gopnik, 2003; Pearl, 1988) and their generalization in probabilistic programs (Goodman et al., 2015) are obvious starting places to look for such a formalism for conceptual structure and higher-order abstractions.

The fact that conceptual knowledge is not a part of our formal semantics does not imply that such knowledge is unrelated to communicating generalizations. In fact, we saw in Experiment 2c that it was participants' predictions about the future that led to the prevalence computation involved in the endorsement decision. Intuitive theories guide these predictions and might provide new insight into classic puzzles of generics. We described in the *Worked Examples* how our model treats "Supreme Court Justices have even social security numbers" as infelicitous even when 100% of justices have even Social Security Numbers, because a speaker's subjective probability that the next justice would have an even Social Security Number is likely 50%. However, were we to learn much more surprising information—for instance, if every Supreme Court Justice in U.S. history had a social security number that was a prime number (a much lower probability outcome)—the shear suspiciousness could compel an observer to revise their theory of the domain (appealing perhaps to a conspiracy), update their subjective probability of future instances, and endorse such a generic.

Not only does conceptual knowledge influence interpretation via beliefs about prevalence, but conceptual knowledge may very well be what a speaker intends to communicate. Like other models in the probabilistic pragmatics tradition, our model distinguishes the relevant variables for the truth-functional semantics (e.g., a threshold and a prevalence) from those that impact the speaker's utility in producing the utterance. A speaker's utility comes from addressing an implicit goal of communication, sometimes referred to as the *Question Under Discussion* or QUD (Roberts, 1996). The separation of the QUD from the literal, truth-functional denotation of the utterance is an important theoretical distinction made explicit in our formal modeling approach (Goodman & Frank, 2016). For simplicity, the model presented here assumes the QUD is about prevalence. However, the model makes nearly identical predictions if the QUD concerns abstract parameters of the prevalence distributions (e.g., if the speaker intends to convey that "K is the kind of thing that Fs"; cf., Prasada & Dillingham, 2006), something much closer to communicating conceptual understanding. This observation may prove useful for constructing models of speakers whose goals are to convey abstract relations between kinds and properties.

Conceptual knowledge concerning the strikingness or dangerousness of properties and their relation to generic language has been of particular interest to psychologists and philosophers since Leslie (2007). We described before that the strikingness of certain properties plausibly influences prevalence priors or a speaker's predictions about future prevalence. We have collected pilot data suggesting these constructs are indeed altered as a result of introducing information about the dangerousness of features to participants (e.g., as in Cimpian et al., 2010). More empirical work is needed to understand whether and how strikingness influences generic understanding above and beyond the probabilistic constructs we have posited in our theory.

The Comparison Class

The probabilistic communicative model we introduce in this article assumes shared background knowledge about the statistics of an event or property in question, represented by the prior belief distribution over the prevalence $P(p)$. We constructed prevalence priors for a property, event, or cause by considering other possible categories having the property, people doing the action, or causes producing the effect, respectively. In Experiment 3, we empirically demonstrated the influence of other categories on endorsements of causal generalizations. Collectively, these other kinds, people, or causes form *comparison classes* against which the referent-category is evaluated. Thus, the prevalence prior $P(p)$ is actually a conditional distribution constructed with respect to some comparison class C: $P(p \mid C)$.

In this article, we have assumed particular values of C for our three case studies (*animals*, *people*, and *possible causes*, respectively). We think the choice of these classes is intuitive, but it is a limitation of this work that we do not derive these choices from more general information-theoretic considerations. The problem of choosing a comparison class, however, is not unique to the language of generalization, but is a problem for any theory of *vague* or *underspecified* language (e.g., gradable adjectives like *tall* and vague quantifiers like *many*; Bale, 2011; Qing & Franke, 2014; Schmidt, Goodman, Barner, & Tenenbaum, 2009; Solt, 2009; Solt & Gotzner, 2012). We have begun to explore how pragmatic reasoning and world knowledge can flexibly adjust the comparison class to appropriately suit the context (e.g., how a "short" basketball player is *short for a basketball player* while a "tall" basketball player is *tall for a person*; Tessler, Lopez-Brau, & Goodman, 2017). It remains to be seen how such principles could operate on richly structured, hierarchical knowledge about categories and properties that would be important for interpreting generic language.

The employment of comparison classes in our model provides additional flexibility toward modeling different communicative goals. The comparison classes used in our case studies were constructed with respect to the category (*other animals*, *other people*, or *other possible causes*). Thus, our endorsement model assumes that the QUD was "*What* has this feature?". Any part of a sentence, however, can be brought into focus (e.g., by prosody) and turned into an answer to a QUD: "*Dogs* have four legs" (as opposed to other animals), "Dogs *have* four legs" (as opposed to eating or doing other things with four legs), "Dogs have *four* legs" (as opposed to two legs, three legs, or six legs), "Dogs have four *legs*" (as opposed to having four of other limbs/body parts), "Dogs *have four legs*" (as opposed to having other features). It has long been noted that the same generalization can be used to address multiple QUDs (Krifka, 1995), and we posit that differences in interpretation are the result of multiple distinct comparison classes competing for influence. Prosody provides a good cue toward resolving the comparison class, and pragmatic reasoning is likely relevant here as well (Declerck, 1986; Tessler et al., 2017). In this work, we focused on category-wise comparison classes purely for methodological convenience. Future work should investigate the factors that give rise to feature-wise interpretations of generalizations (e.g., "Dogs *have four legs*," as opposed to having other features), other QUDs generics can address, how the relevant comparison classes are constructed, and how these inferences interact with threshold inference.

Genericity and Vagueness

We have argued that utterances that communicate generalizations are vague descriptions of prevalence in a way analogous to how gradable adjectives like *tall* are vague descriptions of some underlying degree scale (e.g., height). Vague predicates like *tall*, however, exhibit a number of additional phenomena that we do not suppose generics or generalizations would necessarily exhibit. For example, vague predicates admit borderline cases (e.g., a person who is neither tall nor not tall). *A priori*, it is unclear whether one could construct borderline cases with generics or habituals without sounding like a contradiction (e.g., "John both runs and doesn't run"). Additionally, sorites paradoxes can be constructed out of arguments using vague predicates, and it is not clear that generics or habituals could be used to construct such arguments (e.g., a person who runs one day a week less than John [who runs habitually] is still a person who runs). No experimental data yet bears on these ideas, but would be useful for further understanding the relationship between the language of generalization and vague predicates.

For now, we understand both genericity and vagueness as analogous linguistic phenomena insofar as they can both be thought of as the result of a listener being uncertainty as to the exact truth conditions (e.g., an uncertain threshold). This sort of "lexical uncertainty" is a kind of *parameter learning* problem (e.g., the listener knows the form of the truth conditions, but not the values of some variables) that can be formally distinguished from other sorts of uncertain truth conditions better modeled as a *structure learning* problem (e.g., lexical ambiguity). Further work should be done to better understand the kinds of context-sensitive linguistic phenomena that are modeled as parameter versus structure learning problems to construct a richer typology of underspecified linguistic expressions.

Our view of understanding the language of generalization involving reasoning about an uncertain threshold is reminiscent of Williamson and Simons (1992)'s epistemicist view on vagueness. This view holds that when a speaker utters a vague statement, there exists an objective set of criteria determining whether the entity in question satisfies the vague predicate (i.e., whether or not the utterance is true). Vagueness emerges from uncertainty about the details of those objective criteria (e.g., the precise value of a threshold). This uncertainty in turn could be the result of a population of speakers using slightly different criteria. Listeners then will not know exactly what kind of speaker they are dealing with *a priori* and normatively should maintain uncertainty about the truth conditions (see Lassiter & Goodman, 2015 for an extended discussion of vagueness involving resolving the value of a free semantic variable in context).

Acquiring the Language of Generalization

Perhaps the most surprising aspect of the language of generalization is how difficult it is to formalize, given how common it is in child-directed and child-produced speech (Gelman, Coley, Rosengren, Hartman, & Pappas, 1998; Gelman, Taylor, Nguyen, Leaper, & Bigler, 2004). Generics are often contrasted with quantifier language (e.g., "some," "most," and "all"), whose truth conditions are easy to formalize but which pose difficulty for young children to acquire (Brandone, Gelman, & Hedglen, 2014; Gelman et al., 2015). Leslie (2008) argues that the necessary complexity of a formal account of generics and the simplicity with which young children acquire generics imply that the normal tools for describing the semantics of quantified utterances (i.e., a truth-functional threshold) are inappropriate for generics.[25]

Rejecting the tools of truth-functional semantics for the language of generalization would be throwing the baby out with the bathwater: In fact, from a learning perspective, our semantics of generics can be viewed as simpler than that of quantifier semantics. With some reasonable assumptions about the hypothesis space of semantic meanings, the acquisition of a threshold-based truth-functional meaning $[\![u]\!](p, \theta){:} = p > \theta$ requires learning three distinct aspects of meaning: (a) the *dimension* being described (i.e., prevalence p), (b) the *polarity* of the relation (i.e., $>$ vs. $<$), and (c) the value of the *threshold* (e.g., $\theta = 0$ for "some," $\theta = 0.5$ for "most").[26] If a learner first acquires the dimension and the polarity (i.e., generics have to do with prevalence in some positive way), a rational learner should then represent uncertainty over possible thresholds θ. For quantifier semantics, a learner would then need to learn the context-invariant value of the threshold θ, but for generics, she could have an adult-like semantics for generics by maintaining uncertainty about the threshold. That is, the language of generalization is learned once aspects (a) and (b) are understood. Thus, under our framework, generics should not only be learned before quantifiers but could also facilitate the acquisition of quantifiers because of their shared logical form.

There is a secondary argument for why our semantics presents an easy learning problem. We have assumed the truth functional threshold θ comes from a uniform distribution over the unit interval [0, 1], which is mathematically equivalent to a *continuously valued* or *soft semantics* wherein the degree to which the utterance is true is proportional to the degree itself, in this case prevalence p: $\int_0^1 \delta_{p>\theta}\, d\theta = p$. The model for generic interpretation (Eq. 1) then becomes: $L(p \mid u) \propto p \cdot P(p)$ This *soft semantics* (corresponding intuitively to a meaning like "the higher the prevalence, the better," or simply "more is better") is perhaps the simplest quantitative semantics one could posit. The difficulty in acquiring the meaning of quantifiers, then, is a difficulty in recognizing fixed-threshold semantics as a special case of this *more is better* semantics. We

[25] This argument is a purely semantic argument that ignores evidence that children's difficulty with quantifiers stems from pragmatic issues (Musolino & Lidz, 2006).

[26] This particular framing of the acquisition process ignores the potentially different impact of pragmatic reasoning on generics versus quantifiers.

leave for future work the precise implementation of such an acquisition model.

Conclusion

It might seem paradoxical that a part of language that is so common in communication and central to learning should be vague. Should not speakers and teachers want to express their ideas as crisply as possible? To the contrary, underspecification can be efficient, given that context can be relied upon to resolve uncertainty in the moment (Piantadosi, Tily, & Gibson, 2012). In our work, context takes the form of shared beliefs between speakers and listeners. By leveraging this common ground, the language of generalizations provides a powerful way to communicate and learn abstract knowledge, which would otherwise be difficult or costly information to acquire through direct experience.

Categories are inherently unobservable. You cannot see the category *dog*, only some number of instances of it. Yet, we easily talk about these abstractions, conveying hard-won generalizations to each other and down through generations. The theory presented here provides the first computational perspective on how we communicate generalizations, illustrating how beliefs play a central role in understanding the meaning of words.

References

Asher, N., & Morreau, M. (1995). What some generic sentences mean. In G. N. Carlson & F. J. Pelletier (Eds.), *The generic book* (pp. 300–338). Illinois: University of Chicago Press.

Bale, A. C. (2011). Scales and comparison classes. *Natural Language Semantics, 19,* 169–190.

Behrens, L. (2005). Genericity from a cross-linguistic perspective. *Linguistics, 43,* 275–344.

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, 39,* 711–738. http://dx.doi.org/10.1111/cogs.12176

Carlson, G. N. (1977). *Reference to kinds in English* (Doctoral dissertation). University of Massachusetts, Amherst, MA.

Carlson, G. N. (1995). Truth conditions of generic sentences: Two contrasting views. In G. N. Carlson & F. J. Pelletier (Eds.), *The generic book* (pp. 224–38). Illinois: University of Chicago Press.

Carlson, G. N. (2005). Generics, habituals and iteratives. In A. Barber (Ed.), *The encyclopedia of language and linguistics* (2nd ed.). Amsterdam, the Netherlands: Elsevier.

Carlson, G. N., & Pelletier, F. J. (1995). *The generic book*. Illinois: Chicago University Press.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367–405. Retrieved from http://psycnet.apa.org/fulltext/1997-03612-007.pdf

Cheng, P. W., & Npvick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99,* 365–382. Retrieved from http://psycnet.apa.org/fulltext/1992-26096-001.pdf

Cimpian, A., Arce, H.-M. C., Dweck, C. S., & Markman, E. M. (2007). Subtle linguistic cues affect children's motivation. *Psychological Science, 18,* 314–316.

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

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

Clark, H. H. (1996). *Using language*. New York, NY: Cambridge University Press.

Cohen, A. (1999). Generics, frequency adverbs, and probability. *Linguistics and Philosophy, 22,* 221–253.

Cohen, A. (2004). Generics and mental representations. *Linguistics and Philosophy, 27,* 529–556. http://dx.doi.org/10.1023/B:LING.0000033851.25870.3e

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology Learning, Memory, and Cognition, 32,* 643–658. http://dx.doi.org/10.1037/0278-7393.32.4.643

Dahl, Ö. (1975). On generics. In E. Keenan (Ed.), *Formal semantics of natural language* (pp. 99–111). United Kingdom: Cambridge University Press.

Declerck, R. (1986). The manifold interpretations of generic sentences. *Lingua, 68,* 149–188.

Declerck, R. (1991). The origins of genericity. *Linguistics, 29,* 79–102.

Degen, J., & Goodman, N. D. (2014). Lost your marbles? The puzzle of dependent measures in experimental pragmatics. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the thirty-sixth annual conference of the Cognitive Science Society*. Quebec City, Quebec, Canada.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336,* 998.

Franke, M. (2014). Typical use of quantifiers: A probabilistic speaker model. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 1–6). Retrieved from https://escholarship.org/uc/item/10z1z670

Gelman, S. A. (2003). *Essential child: Origins of essentialntialism in everyday thought*. New York, NY: Oxford University Press.

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

Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual Review of Psychology, 60,* 115–140. http://dx.doi.org/10.1146/annurev.psych.59.103006.093659.LEARNING

Gelman, S. A., Coley, J. D., Rosengren, K. S., Hartman, E., & Pappas, A. (1998). Beyond labeling: The role of maternal input in the acquisition of richly structured categories. *Monographs of the Society for Research in Child Development, 63,* i. http://dx.doi.org/10.2307/1166211

Gelman, S. A., Goetz, P. J., Sarnecka, B. W., & Flukes, J. (2008). Generic language in parent-child conversations. *Language Learning and Development, 4,* 1–31. http://dx.doi.org/10.1080/15475440701542625.Generic

Gelman, S. A., Leslie, S.-J., Was, A. M., & Koch, C. M. (2015). Children's interpretations of general quantifiers, specific quantifiers and generics. *Language, Cognition and Neuroscience, 30,* 448–461. http://dx.doi.org/10.1080/23273798.2014.931591

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

Gelman, S. A., Taylor, M. G., Nguyen, S. P., Leaper, C., & Bigler, R. S. (2004). Mother-child conversations about gender: Understanding the acquisition of essentialist beliefs. *Monographs of the Society for Research in Child Development, 69,* vii–127. http://dx.doi.org/10.1111/j.1540-5834.2004.06901001.x

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20,* 818–829.

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 ed.). Hoboken, NJ: Wiley-Blackwell.

Goodman, N. D., & Stuhlmüller, A. (2014). *The design and implementation of probabilistic programming languages*. Retrieved from http://dippl.org

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Lawrence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–653). Cambridge, MA: MIT Press.

Gopnik, A. (2003). The theory theory as an alternative to the innateness hypothesis. In L. M. Antony (Ed.), *Chomsky and his critics* (pp. 238–254). New York, NY: Basil Blackwell.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Ed.), *Syntax and semantics* (Vol. 3, pp. 41–58). New York, NY: Academic Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384. http://dx.doi.org/10.1016/j.cogpsych.2005.05.004

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116,* 661–716. http://dx.doi.org/10.1037/a0017201

Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. New Jersey: Princeton University Press.

Herbelot, A., & Copestake, A. (2011). Formalising and specifying underquantification quantification resolution. In J. Bos & S. Pulman (Eds.), *Proceedings of the Ninth International Conference on Computational Semantics (IWCS)* (pp. 165–174). Oxford, United Kingdom: Association for Computational Linguistics.

Hume, David (1738). *A treatise of human nature*. United Kingdom: Oxford University Press.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition, 161,* 80–93. http://dx.doi.org/10.1016/j.cognition.2017.01.010

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science, 10,* 307–321. http://dx.doi.org/10.1111/j.1467-7687.2007.00585.x

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy, 30,* 1–35. http://dx.doi.org/10.1007/s10988-006-9008-0

Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2012). Inferences about members of kinds: The generics hypothesis. *Language and Cognitive Processes, 27,* 887–900. http://dx.doi.org/10.1080/01690965.2011.601900

Krifka, M. (1995). Focus and the interpretation of generic sentences. In G. N. Carlson & F. J. Pelletier (Eds.), *The generic book* (pp. 238–264). Illinois: University of Chicago Press.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. *Proceedings of SALT, 23,* 587–610.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese, 194,* 3801–3836.

Leslie, S.-J. (2007). Generics and the structure of the mind. *Philosophical Perspectives, 21,* 375–403.

Leslie, S.-J. (2008). Generics: Cognition and acquisition. *Philosophical Review, 117,* 1–47.

Leslie, S.-J., & Gelman, S. A. (2012). Quantified statements are recalled as generics: Evidence from preschool children and adults. *Cognitive Psychology, 64,* 186–214. http://dx.doi.org/10.1016/j.cogpsych.2011.12.001

Levinson, S. C. (1995). Interactional biases in human thinking. In E. N. Goody (Ed.), *Social intelligence and interaction* (pp. 221–260). New York, NY: Cambridge University Press.

Liebesman, D. (2011). Simple generics. *Noûs, 45,* 409–442.

McGuire, W. J., & McGuire, C. V. (1986). Differences in conceptualizing self versus conceptualizing other people as manifested in contrasting verb types used in natural speech. *Journal of Personality and Social Psychology, 51,* 1135–1143. http://dx.doi.org/10.1037/0022-3514.51.6.1135

Montague, R. (1973). The proper treatment of quantification in ordinary English. In J. Kulas, J. H. Fetzer, & T. L. Rankin (Eds.), *Philosophy, language, and artificial intelligence* (pp. 141–162). Dordrecht, the Netherlands: Springer. Retrieved from http://semantics.uchicago.edu/kennedy/classes/s08/semantics2/montague73.pdf

Musolino, J., & Lidz, J. (2006). Why children aren't universally successful with quantification. *Linguistics, 44,* 817–852.

Nickel, B. (2008). Generics and the ways of normality. *Linguistics and Philosophy, 31,* 629–648. http://dx.doi.org/10.1007/s10988-008-9049-7

Nickel, B. (2016). *Between logic and the world*. United Kingdom: Oxford 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,* 339–363. http://dx.doi.org/10.1037/0033-295X.90.4.339

Orvell, A., Kross, E., & Gelman, S. A. (2017). How "you" makes meaning. *Science, 355,* 1299–1302. Retrieved from http://science.sciencemag.org/content/355/6331/1299.full

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible reasoning*. Los Altos, CA: Morgan Kaufmann Publishers.

Pelletier, F. J., & Asher, N. (1997). Generics and defaults. In J. van Benthem & A. Ter Meulen (Eds.), *Handbook of logic and language* (pp. 1125–1177). Cambridge, MA: The MIT Press.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition, 122,* 280–291. http://dx.doi.org/10.1016/j.cognition.2011.10.004

Prasada, S. (2000). Acquiring generic knowledge. *Trends in Cognitive Sciences, 4,* 66–72. http://dx.doi.org/10.1016/S1364-6613(99)01429-1

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition, 99,* 73–112. http://dx.doi.org/10.1016/j.cognition.2005.01.003

Prasada, S., Hennefield, L., & Otap, D. (2012). Conceptual and linguistic representations of kinds and classes. *Cognitive Science, 36,* 1224–1250. http://dx.doi.org/10.1111/j.1551-6709.2012.01254.x

Prasada, S., Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2013). Conceptual distinctions amongst generics. *Cognition, 126,* 405–422. http://dx.doi.org/10.1016/j.cognition.2012.11.010

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. *Proceeding of SALT, 24,* 587–610.

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology, 33,* 12–21. http://dx.doi.org/10.1037/0012-1649.33.1.12

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences of the United States of America, 109,* 13526–13531. http://dx.doi.org/10.1073/pnas.1208951109

Ritchie, D., Stuhlm̈uller, A., & Goodman, N. D. (2016). C3: Lightweight incrementalized MCMC for probabilistic programs using continuations and callsite caching. In A. Gretton & C. C. Robert (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Retrieved from http://arxiv.org/abs/1509.02151

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. *OSU Working papers in linguistics* (pp. 91–136). Columbus: Department of Linguistics, Ohio State University.

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,* 463–481.

Rothbart, M., Fulero, S., Jensen, C., Howard, J., & Birrell, P. (1978). From individual to group impressions: Availability heuristics in stereotype formation. *Journal of Experimental Social Psychology, 14,* 237–255. http://dx.doi.org/10.1016/0022-1031(78)90013-6

Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How tall is tall? Compositionality, statistics, and gradable adjectives. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 2759–2764). Amsterdam, the Netherlands: Cognitive Science Society.

Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did she jump because she was the big sister or because the trampoline was safe? Causal inference and the development of social attribution. *Child Development, 84,* 443–454. http://dx.doi.org/10.1111/j.1467-8624.2012.01865.x

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323. http://dx.doi.org/10.1126/science.3629243

Shipley, E. F. (1993). Categories, hierarchies, and induction. *The Psychology of Learning and Motivation, 30,* 265–301.

Solt, S. (2009). Notes on the comparison class. In R. Nouwen, R. van Rooij, U. Sauerland, & H. C. Schmitz (Eds.), *International workshop on vagueness in communication* (pp. 189–206). Berlin, Germany: Springer.

Solt, S., & Gotzner, N. (2012). Experimenting with degree. *Proceedings of SALT, 22,* 166–187.

Sterken, R. K. (2015). Generics in context. *Philosophers' Imprint, 15,* 1–30.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–640. http://dx.doi.org/10.1017/S0140525X01000061

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331,* 1279–1285. http://dx.doi.org/10.1126/science.1192788

Tessler, M. H., & Franke, M. (2018). Not unreasonable: Carving vague dimensions with contraries and contradictions. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. Madison, WI.

Tessler, M. H., Lopez-Brau, M., & Goodman, N. D. (2017). Warm (for winter): Comparison class understanding in vague language. In *Proceedings of the 39th annual conference of the cognitive science society*. London, United Kingdom.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185,* 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

Williamson, T., & Simons, P. (1992). Vagueness and ignorance. *Proceedings of the Aristotelian Society, Supplementary Volumes, 66,* 145–177.

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

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 5012–5015. http://dx.doi.org/10.1073/pnas.0704450105

(*Appendices follow*)

Appendix A

The Relationship Between the Prevalence Prior and Cue Validity

Cue validity is defined for a particular category–property pair (e.g., *mosquitos* and *carry malaria*), and relates to the referent prevalence (e.g., how many mosquitos carry malaria) via Bayes' Rule:

$$P(k|f) = \frac{P(f|k) \cdot P(k)}{Z}$$

where $Z = \sum_{k' \in K} P(f \mid k') \cdot P(k')$, the average or marginal prevalence of the feature (e.g., *carrying malaria*) in other categories k'.

The prevalence prior $P(p)$ used in the generic interpretation model (Eq. 1) is a probability distribution over prevalence for different categories k'.

Claim: The normalizing constant for computing cue validity is equal to the expected value of the prevalence prior distribution: $\mathbb{E}[P(p)] = Z$

Proof: By the definition of the expectation of a distribution:

$$\mathbb{E}[P(p)] = \sum_p p \times P(p) \tag{5}$$

The probability of a prevalence p can be decomposed into the prior probability of a category k and the likelihood of the prevalence p given that category k: $P(p) = P(p \mid k) \cdot P(k)$. We assume here, without loss of generality, that each category corresponds to one and only one prevalence p. Thus, $P(p \mid k) = 1$ if $k \in K_p$, a set of categories that have a given prevalence: $K_p = \{k' : p_{k'} = p\}$. Then, consider the partition of the set of all categories K into nonoverlapping K_p. Thus:

$$\begin{aligned} P(p) &= \sum_{k' \in K_p} P(p|k') \cdot P(k') \\ &= \sum_{k' \in K_p} P(k') \end{aligned} \tag{6}$$

since $\forall k' \in K_p$: $P(p \mid k') = 1$. Returning to Eq. 5, we have:

$$\begin{aligned} \mathbb{E}[P(p)] &= \sum_p p \sum_{k' \in K_p} P(k') \\ &= \sum_p \sum_{k' \in K_p} p \cdot P(k') \end{aligned} \tag{7}$$

The set of all partitioned subsets K_p is in a one-to-one correspondence with the set of all prevalences p. Thus, we have:

$$= \sum_{K_p} \sum_{k' \in K_p} p \cdot P(k') \tag{8}$$

Then, since $\cup_p K_p = K$, we have

$$\begin{aligned} &= \sum_{k' \in K} p \cdot P(k') \\ &= \sum_{k' \in K} P(f|k') \cdot P(k') \\ &= Z \end{aligned} \tag{9}$$

Appendix B

Measuring Cue Validity

In Experiment 1, we articulated an alternative model by measuring cue validity (and prevalence) and predicting generic endorsement from a linear combination of these parameters. In a small review of the literature, we discovered different methods for measuring cue validity; in piloting, we found these different methods led to different results. Therefore, we propose three *a priori* desiderata that a measurement of cue validity should satisfy. We describe two experiments that represent the primary methods for measuring cue validity and compare them with these desiderata in mind. Finally, we compare the cue validity measured using these different methods to cue validity derived from our prevalence prior elicitation task (Experiment 1b, main text).

Desiderata

Measuring cue validity involves collecting participants' judgments that relate to the probability that an exemplar is a member of a kind given that it has a feature: $P(x \in k \mid x \in f)$. There are several ways one could measure cue validity. Here we consider two measures: The first has participants estimate the cue validity probability $P(x \in k \mid x \in f)$ directly, a common technique in the literature on generic language (e.g., Khemlani et al., 2012; Prasada et al., 2013), and another has participants freely produce categories given a feature (i.e., draw a sample from the conditional distribution on kinds given a feature; Cree et al., 2006). We will refer to the former as the *direct question* method and the latter as the *free production* method.

Are the *direct question* and the *free productions* equally valid for measuring cue validity? We propose *a priori* three boundary conditions that a measurement of cue validity should satisfy. For each condition we provide four examples from our larger stimulus set on generics (Case Study 1) that will be used to evaluate each measure.

(*Appendices continue*)

1. Completely diagnostic features: Some features are only present in one (or a very small number) of categories. Examples include: *carrying malaria* (mosquitos), *carrying Lyme disease* (ticks, deer), *having manes* (lions), and *having pouches* (marsupials, including most famously: kangaroos). The cue validity of these features for the corresponding categories should be high (at least 0.5 and possibly close to 1).

2. Completely absent features: Many features are completely absent in many kinds. For these, the cue validity should be extremely low or 0. There are infinite examples. The ones we will use are *has wings* (leopard), *has a mane* (shark), *has spots* (kangaroo), and *has a pouch* (tiger).

3. Completely undiagnostic features: A number of features are shared by almost every category. The cue validity of these features for particular categories should be extremely low or 0. The ones we will use are: *is female* (robin), *is male* (lion), *is juvenile* (kangaroo), and *is full-grown* (leopard). Learning that an entity is female tells you almost nothing about what kind of animal it is.

We collected cue validity ratings by running both a direct question and a free production experiment. For the free production experiment, the cue validity is the proportion of responses of the target category (e.g., *mosquitos*) for the property (e.g., *carries malaria*). Of primary interest is the measurement for the desiderata items described above. Links to the experiments can be found on https://github.com/mhtess/genlang-paper.

Experiments

Materials were the same for both experiments. These were a collection of familiar properties and animal categories used in Experiment 1a (endorsement of generic statements) described in the main text. There were 21 properties in total.

Direct Question Experiment

Participants. We recruited 40 participants from Amazon's MTurk. Participants were restricted to those with U.S. IP addresses and who had at least a 95% work approval rating. The experiment took on average 5 min and participants were compensated $0.75 for their work.

Procedure. Following the procedure in Khemlani et al. (2012) and Prasada et al. (2013), participants were presented with prompts of the following form:

> Imagine you come across a thing that F. what are the odds that it is a K?

Participants responded using a slider bar with endpoints labeled "unlikely" and "likely." The slider appeared with no handle present; participants had to click on the slider for the slider handle to appear.

Participants completed the 30 target trials (corresponding to the 30 generic statements used in Experiment 1a) in addition to ten filler trials (total number of trials = 40). The filler trials were made up of random category–property pairings. Trials were presented in a randomized order.

Free Production Experiment

Participants. We recruited 100 participants from Amazon's MTurk. Participants were restricted to those with U.S. IP addresses and who had at least a 95% work approval rating. The experiment took on average 3 minutes and participants were compensated $0.40 for their work.

Procedure. On each trial, participants were presented with prompts of the following form:

> Imagine you come across a thing (animal or insect) that F. What do you think it is?

Participants responded by filling in a text box with their response for 21 trials in total, one for each property. No filler trials were used. Trials were presented in a randomized order.

Free production data processing. To process the free production, we forced all characters in a response to lower case, removed spaces, and made all terms into singular terms (e.g., "lions" –> "lion"). As well, "mosquito" was a commonly misspelled label; we counted anything that started with "mosque," "mesqu," "misqu," "mosiq," as "mosquito."

To calculate confidence intervals for the free production data, we resampled participants (with replacement) and computed the proportion of responses that were of the target category (e.g., the proportion of *mosquito* responses for the cue *carries malaria*). We did this 1,000 times to generate an empirical distribution from which 95% CIs could be calculated.

Results and Evaluation

We are interested in the results of each measure (direct question and free production) for the three conditions corresponding to the desiderata outlined above. To evaluate each measure, we selected four example property–category pairs that we believe are unambiguous instances of the boundary conditions described above (these items are described above with the desiderata).

(*Appendices continue*)

Figure B1A shows the results for the twelve items of interest for both measurements. We see that for the *false features*, both measures behave as desired (hypothesized results shown by the dotted line): The cue validity of a feature that is not present in the category is zero or near-zero. For *diagnostic features*, both measures also behave reasonably: Learning that an entity has malaria strongly implies that it is a mosquito. However, there is some evidence that the free production measurement is more sensitive than the direct-question measure. *Having a mane* is strongly diagnostic for a *lion* but also for a *horse* (and so the overall cue validity of having a mane for a lion is around 0.5). *Carrying Lyme disease* is mostly diagnostic for a *tick* but also *deer* (and, thus, the cue validity for tick is not maximal). These subtle differences among diagnostic features are picked up by the free-production measure but not by the direct-question measure.

The free production and direct question measures deviate most strongly in their characterization of the undiagnostic features. Learning that an entity is female should not imply that it is a robin,

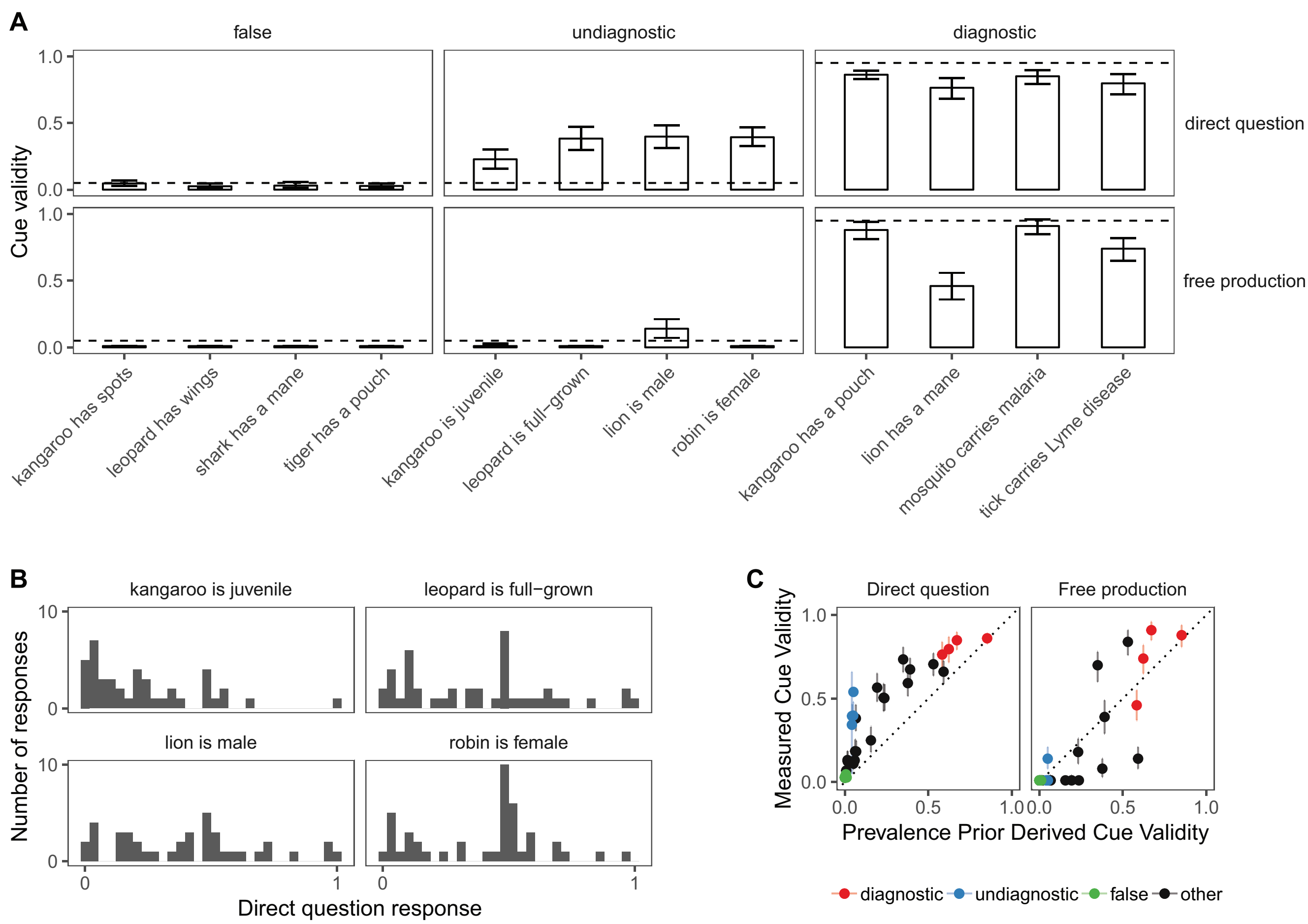

Figure B1. (A) Empirically measured cue validity for two different tasks. Items are grouped by whether the property is never present in the category (false), the property is always present in the category and every other category (undiagnostic), or present in the category and absent from most other categories (diagnostic). Dotted lines denote theoretical cue validity representing the desiderata (see text). (B) Raw empirical distributions for the undiagnostic features in the direct question task. (C) Correspondence between measured cue validity and prevalence prior derived cue validity. See the online article for the color version of this figure.

(*Appendices continue*)

which is accurately reflected in the free production measure but not in the direct question measure. Figure B1B shows the raw empirical distributions of responses for the direct question measure for undiagnositic features. We observe that participants respond to this question for undiagnostic features in one of two ways: (1) reporting near-0 likelihood (hypothesized response) or (2) reported near-0.5 likelihood. This latter response option may reflect participants "opting out" of a response (e.g., signaling "I don't know"). For example, in response to the question "There is a thing that is female. What are the odds that it is a robin?", a person could say they have no evidence to suggest that it is, besides the very fact that the experimenter asked the question. Participants may cope with the awkwardness of the question by placing the slider bar in the middle of the scale.

Comparison with Prevalence Prior Derived Cue Validity

To further understand these measures of cue validity, we compare them to cue validity derived from our prevalence prior elicitation task (Experiment 1b). Experiment 1b is not perfectly designed for this comparison, as we supplied participants with half of the animal categories that they rated (the other half was freely generated by participants); including these supplied categories was important to measure the referent-prevalence of interest (e.g., the percentage of mosquitos that carry malaria). Including them in this analysis, however, potentially distorts the prior probability of categories $P(k)$.

In this analysis, we treat each category entry from Experiment 1b (participant free production or experimentally supplied category like mosquitos) as contributing to the prevalence prior. This results in the prevalence prior favoring kinds that are easy to produce (like dogs and cats; plausibly a good approximation for $P(k)$) as well as favoring the experimentally supplied kinds (like mosquitos and robins).

We compute cue validity from the prevalence prior using Bayes' rule. Figure B1C shows the two measurements of cue validity as they relate to the prevalence prior derived cue validity. Across the 30 property–category pairs, the prevalence prior derived cue validity is highly associated with both measurements: $r^2_{direct}(30) = 0.782$; $MSE_{direct} = 0.0481$ and $r^2_{free}(30) = 0.739$; $MSE_{free} = 0.0244$. Of primary interest is how the measurements behave for desiderata items. We see that the prevalence prior derived cue validity converges with the free production measurement for the desiderata items ($r^2_{free}(12) = 0.947$; $MSE_{free} = 0.00822$), whereas the direct question measurement overestimates the cue validity of undiagnositic features ($r^2_{direct}(12) = 0.779$; $MSE_{direct} = 0.0561$).

The points of largest deviation for the free-production measurement from the prevalence prior derived measurement occur where the prevalence prior derived measure rates the cue validity as relatively high when the free-production measure gives the item low cue validity (two black points in Figure 12C, high X-value, low Y-value). These two items are: (*is red*, *cardinal*) and (*is white*, *swan*). These items should have relatively low cue validity and are overestimated by the prevalence prior because of the prior on categories $P(k)$ overrepresents the categories that were supplied to every participant and, thus, get a higher weighting in the prior for deriving cue validity.[27] In comparison, the direct question measurement overestimates cue validity for almost all of the items relative to the cue validity derived from the prevalence prior.

Summary

Cue validity is a commonly used measurement for understanding generic truth judgments (e.g., Khemlani et al., 2012; Prasada et al., 2013). We observed different measurements used in the literature and articulated three *a priori* desiderata to validate a measure of cue validity. We found that the *free production* measurement (i.e., participants freely produced categories given a feature), and not the direct question measurement (i.e., participants provide a likelihood judgment of a particular category given the feature), satisfied all three boundary conditions. In addition, cue validity derived from our prevalence prior measurement (Experiment 1b) also satisfied these boundary conditions. Researchers interested in comparing cue validity to generic truth judgments should use a free production paradigm for measuring cue validity.

[27] This deviation could be reduced in future experiments by performing a fully free production version of the prevalence prior task (i.e., without supplying the referent categories).

(*Appendices continue*)

Appendix C

Bayesian Data Analysis

In our three case studies, we compare an information-theoretic, computational model of endorsement to human endorsements of generalizations in language. The model has a single free parameter: the optimality parameter λ in Eq. 3. We analyze this model using a Bayesian data analytic approach, where we jointly infer the value of this single model parameter λ together with parameters that govern the prevalence priors $P(p)$ (Eq. 1) and referent prevalence p (Eq. 3) for each item. To pin down the prevalence prior and referent prevalence parameters, we use the data directly related to those parameters (e.g., prior elicitation data). Incorporating all data sources into a single Bayesian data analysis model is the appropriate way to track measurement uncertainty for all measurements simultaneously. In this appendix, we describe this procedure in more detail for each case study.

Modeling Prevalence Priors

In Case Study 1: Generic Language (Experiment 1b), we elicited the prevalence prior by asking participants about the prevalence of features for individual categories. We performed an analogous elicitation in Case Study 3: Causal Language (Experiment 3a). We describe the analysis using generics as our running example, but a parallel analysis was done for causal language.

Participants' responses in the prior elicitation task can be thought of as samples from the prevalence prior distribution. Formally, we assume the prior data (analyzed independently for each property) was generated from one of two underlying distributions: a distribution corresponding to those kinds with a stable causal mechanism that *could* give rise to the property ($\mathcal{D}_{stable}$) and a "transient cause" distribution corresponding to those kinds without a stable mechanism ($\mathcal{D}_{transient}$). The "transient" distribution intuitively corresponds to categories that do not have the feature normally, but potentially could acquire the feature by accidental forces (e.g., a lion, who through some genetic mutation, reproduces by laying eggs). We model this distribution as a Beta distribution that heavily favors probabilities near 0: $\text{Beta}(\gamma = 0.01, \delta = 100)$.[28] The "stable" distribution is modeled as a Beta distribution with unknown parameters $\text{Beta}(\gamma, \xi)$.[29] Finally, we assume that these two components combine with mixture weighting ϕ such that the data we observe is

$$P(d) = \phi \cdot \text{Beta}(d|\gamma, \xi) + (1 - \phi) \cdot \text{Beta}(d|\gamma = 0.01, \xi = 100)$$

We put the following priors over the latent parameters of the model:

$$\phi_i \sim \text{Uniform}(0, 1)$$
$$\gamma_i \sim \text{Uniform}(0, 1)$$
$$\xi_i \sim \text{Uniform}(0, 100)$$

where i ranges over the different properties (e.g., *lays eggs*, *carries malaria*).

To learn about the credible values of the parameters, we ran separate MCMC chains for each item, collecting 75,000 samples, removing the first 25,000 for burn-in. To see how well the mixture model fits the prevalence prior data, we use the inferred parameters to generate new data. The data generated from the model's posterior is called the *posterior predictive distribution* and is an important step in model criticism. If the model is a good representation of the data, the posterior predictive data will align with the observed experimental data. We construct a posterior predictive distribution by "forward sampling" the model (i.e., generating new data given the inferred parameter values).[30] Representative posterior predictive results are shown in Figure 5B (main text).

In Case Study 2 on habitual language (Experiment 2a), we asked participants about parameters of this mixture model (by having participants answer questions about different kinds of people) rather than having participants give samples (e.g., by listing their friends and family members, and rating how often they did certain actions). In pilot testing, we found these different methodologies to give comparable results and we opted to ask about hypothetical people to probe about potentially undesirable habits of participants' friends and family (e.g., how often they use cocaine). The questions used in this structured elicitation task are described in the main text.

[28] Note that we use the noncanonical mean γ and concentration ξ (or, inverse-variance) parameterization of the Beta distribution rather than the canonical shape (or pseudocount) parameterization, for ease of posterior inference. The shape parameterization can be recovered using: $\alpha = \gamma \cdot \xi$; $\beta = (1-\gamma) \cdot \xi$.

[29] Because the Beta distribution is not defined at the points 0 and 1, we add ε to the 0 responses and round 1 to 0.99. Similar results can be obtained by rounding 0 responses to 0.01. Alternatively, the "transient" distribution could be defined as a Delta distribution at 0, and 0 responses could remain in their raw form. Adjusting 1 responses to $1 - \varepsilon$ leads to improper inferences for this simple two-component model, as $1 - \varepsilon$ is only likely under a highly left-skewed distribution; treating 1 as $1 - \varepsilon$ disproportionately influences the shape of $\mathcal{D}_{stable}$, forcing it to favor probabilities close to 1. This problem does not appear for 0 being adjusted to ε because the "transient" distribution already expects such low values.

[30] This forward sampling can be described by the following algorithm: First, flip a coin weighted by ϕ. If it comes up heads, we then sample from the "stable" component: $\text{Beta}(\gamma, \xi)$. If it comes up tails, we sample from the "transient" component: Beta(0.01,100). We do this many times using the posterior distribution to generate a distribution over predicted prevalence ratings.

(*Appendices continue*)

Modeling Referent-Prevalence

In Case Study 1 (generics), we used participants prevalence ratings for the category-of-interest in our generic sentences as the referent-prevalence that is used in the endorsement model (Eq. 3). For a given generic sentence (e.g., "Robins lay eggs"), we took the prevalence ratings for the referent-category (e.g., the percentage of robins that lay eggs) from the prior elicitation task (Experiment 1b) and assumed those were generated from a single Beta distribution. We assumed the following priors on the parameters:

$$\gamma_i \sim \text{Uniform}(0, 1)$$
$$\xi_i \sim \text{Uniform}(0, 100)$$

We took samples from the posterior predictive of this Beta distribution (i.e., reconstructed prevalence ratings) as the *referent-prevalence* used in the model.

In Experiment 2b (Habitual endorsement), we used the frequency given to participants in the experimental prompt (e.g., three times in the past week) as the referent-prevalence. In Experiment 2c ("What is prevalence?"), we compared two endorsement models that differed in their representation of referent-prevalence. For one model (past frequency model), the actual frequency given to participants in the experimental prompt was assumed to be the referent prevalence (same as in Experiment 2b); for the other model (predictive frequency model), we used the mean elicited frequency from the *predictive frequency* condition (participants predictions about how often the person would do the action in the next time interval; see main text for details). In Case Study 3 (Causal endorsement), we used the proportion of successful causal events given to participants in the experimental prompt (e.g., 70 out of 100 uses of Herb C made animals sleepy).

Jointly Modeling Referent-Prevalence, Prevalence Priors, and Generic Endorsements

To fit the generic endorsement models, we incorporate them into the Bayesian data analytic model of the prevalence prior data (described above) to create a single, joint-inference model where the optimality parameter λ (Eq. 3) is inferred jointly with all the other latent parameters of the full model (the referent-prevalence p for each category k and property f and the parameters of the prevalence priors $P(p)$ for each property f) using data from Experiments 1a and 1b (Figure C1). For the parameters of the prevalence priors, we use the same priors described in Experiment 1b; for the speaker optimality parameter, we use a prior with a range consistent with the previous literature that uses the same model class: $\lambda \sim \text{Uniform}(0, 5)$. We learn about the *a posteriori* credible values of the joint inference models by collecting samples from 3 MCMC chains of 150,000 iterations removing the first 50,000 iterations for burn-in, using an incrementalized version of the Metropolis-Hastings algorithm (Ritchie et al., 2016). This algorithm is useful for models with many variables that only affect a subset of the full model's predictions (e.g., models with by-item or by-participant parameters, wherein those additional parameters mostly only influence predictions for those items or participants).

Supplementary Model Criticism

In addition to examining the posterior predictive distribution on endorsement judgments (presented in main text), we examined the marginal posteriors on parameters of the prevalence priors and referent-prevalence. These marginal distributions are important to examine to confirm that they have not changed substantially from the parameters inferred from their primary data sources in isolation. For example, when modeling the referent-prevalence data in isolation, the model infers that roughly 65% of robins lay eggs, as that is what participants on average produce in the prevalence elicitation task.[31] If the joint inference model (which models all data sources—referent prevalence, prevalence prior, and generic endorsement—simultaneously) infers referent-prevalence values substantially different from those inferred by a model of referent prevalence in isolation, that would suggest that the joint-model is distorting the prevalence parameters to accommodate the endorsement data. Such a result would call into question the inferences we as scientists derive from the joint inference model. For example, incorporating a linear regression model (of the kind presented as alternative models in the main text) into this Bayesian joint-inference analysis model produces posterior predictions that match the generic endorsement data surprisingly well (e.g., that model predicts "Robins lay eggs" is true). Such a model is only able to do this, however, by distorting the referent-prevalence data, inferring that 100% of robins lay eggs; thus, the linear model in this joint-inference analysis framework sacrifices its goodness-of-fit to the referent-prevalence data to increase its goodness-of-fit to the endorsement data.[32] Such a distortion manifests as a difference between the inferred parameters given only the referent-prevalence data and given the full joint model (all data sources simultaneously).

To investigate this distortion effect in the parameters, we compare the values inferred for the parameters governing the prevalence priors and referent prevalence variables before and after the joint-inference model sees the generic endorsement data. Specifically, we infer the parameters for prevalence priors and referent prevalence by constructing single models of these tasks and compare the inferred values to those that result from the joint-inference model. The inferred parameters for these two models are shown in Figure C2. We found that the referent-prevalences and prevalence priors inferred under the joint model were almost indistinguishable from those inferred using only the referent-prevalence and prevalence prior data, respectively (numerical results reported in main text). These results confirm that modeling all three data sources simultaneously does not distort some data sources (e.g., referent prevalence) to provide good fits for others (e.g., generic endorsement).

[31] Most participants report that 50% of robins lay eggs, while a minority respond 100%.

[32] This distortion effect is why we account for measurement uncertainty in the linear models by bootstrapping the data that forms their predictors, rather than performing a Bayesian analysis.

(*Appendices continue*)

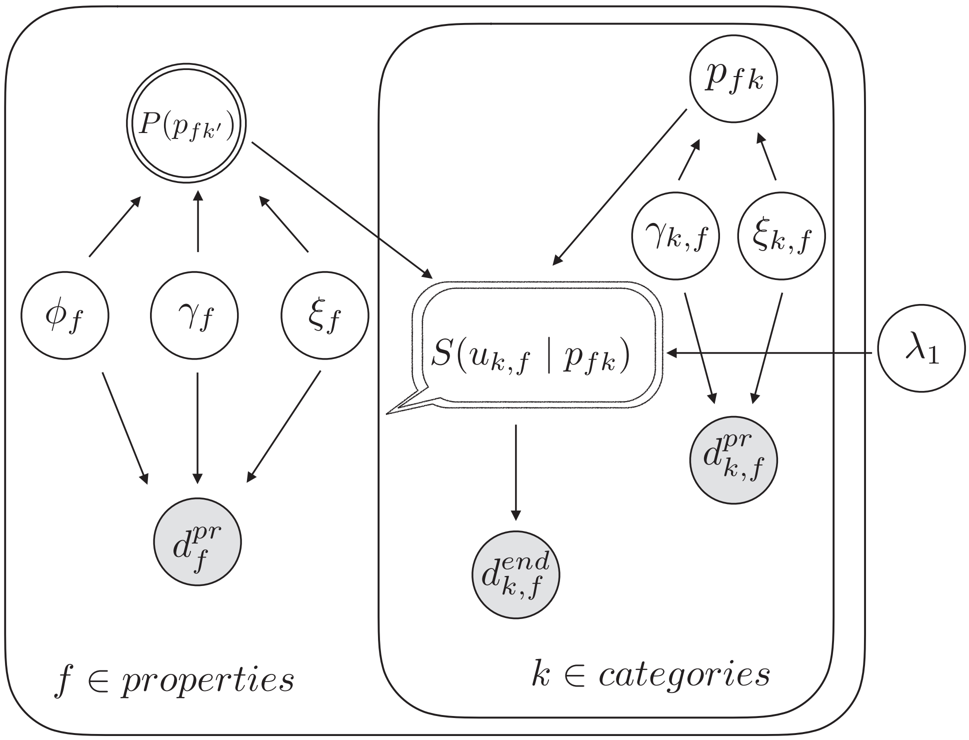

Figure C1. Quasi-graphical model corresponding to the fully Bayesian data analysis of the endorsement model for Case Study 1 (generic language). The prevalence prior data d_f^{pr} is assumed to be generated from the mixture model validated in Experiment 1b, which has three parameters: mean of the stable-cause distribution γ_f and concentration (inverse-variance) of the stable-cause distribution ξ_f and the mixture parameter ϕ_f. The referent-prevalence $d_{k,f}^{pr}$ is generated from a Beta distribution with parameters: mean $\gamma_{k,f}$ and concentration (inverse-variance) $\xi_{k,f}$. The posterior predictives of the prevalence prior $P(p_{fk'})$ and the referent prevalence p_{fk} are then fed into the RSA speaker model *S*, which predicts the generic endorsement data $d_{k,f}^{end}$. The speaker model *S* also takes in the single free parameter λ which operates as a soft-max function. This overall structured is repeated (except λ) for each of the unique properties *f* and categories *k* that correspond to the generic sentences in our stimulus set. Note that *S* corresponds to a probabilistic function and not a random variable that is standard in graphical model notation; *S* cannot be represented by a graphical model because it has recursion. This entire BDA model is duplicated for the lesioned, fixed-threshold model (that only differs in the definition of *S*). The BDA model for habituals and causals mirrors this one, except they do not infer a referent prevalence p_{fk} (they are assumed to be the same as those experimentally supplied to participants).

(*Appendices continue*)

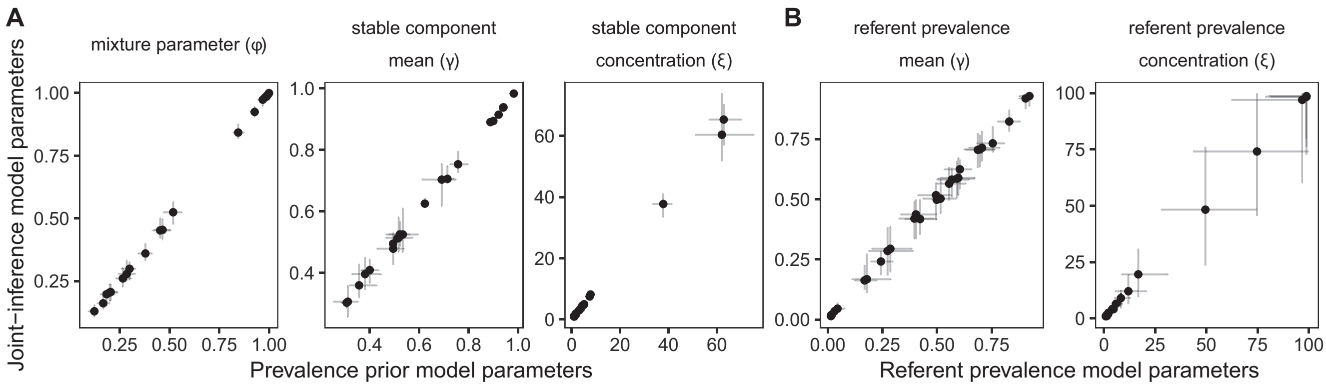

Figure C2. No evidence for parameter distortion caused by a joint-inference model. X-Axis denotes Maximum A-Posteriori (MAP) estimates inferred using only a single data source (A: Prevalence prior data. B: Referent prevalence data). Y-Axis denotes MAP estimates inferred using the joint-inference model which models all data sources simultaneously. Distortions would manifest by systematic deviations from the y = x line.

Appendix D

Habituals Items

Habitual	Referent frequency	Enabling	Preventative
Climbs mountains	5 years*, 2 years, year	Yesterday, William remembered how much fun that was bought a lot of new mountain climbing gear.	Yesterday, William turned 80 and gave up all strenuous physical exercise because his doctor said it would be deadly.
Does cocaine	5 years, year*, month, week	Yesterday, Tim wanted to get high and bought some cocaine.	Yesterday, William realized he is severely allergic to cocaine and will no longer do it.
Drinks beer	year, month*, 2 weeks, week	Yesterday, Tim wanted to get tipsy and bought a six-pack of beer.	Yesterday, William gave up alcohol and entered into Alcoholics Anonymous.
Drinks coffee	year, month*, 2 weeks, week	Yesterday, William wanted a morning jolt and bought a pound of fresh roasted coffee.	Yesterday, Veronica developed a caffeine allergy and decided to give up all caffeine.
Eats caviar	5 years*, year, month	Yesterday, William learned about the dietary benefits of eating caviar and bought a jar at the supermaket.	Yesterday, Tina developed a seafood allergy.
Eats peanut butter	5 years, year*, month	Yesterday, Veronica learned about the dietary benefits of eating peanut butter and bought a jar at the supermaket.	Yesterday, Ted developed a peanut allergy.
Goes to the ballet	2 years, year, month		
Goes to the movies	2 years, year, month		
Hikes	2 years, year*, 2 months, week	Yesterday, William remembered how much fun those times were and bought a lot of new hiking gear.	Yesterday, Vince was in a motorcycle accident and will never walk again.
Listens to live music	year, month, week		
Listens to Pandora	year, month, week		
Plays tennis	5 years, 2 years, year*	Yesterday, Tim remembered how much fun that was and bought a new tennis racket.	Yesterday, William developed crippling arthritis in both elbows and can only move his arms extremely slowly.

(*Appendices continue*)

Appendix D (*continued*)

Habitual	Referent frequency	Enabling	Preventative
Plays the banjo	5 years, 2 years, year*	Yesterday, William remembered how much fun that was and joined his friend & quotechars band as the banjoist.	Yesterday, William developed crippling arthritis in his hands and no longer can play musical instruments.
Runs	2 years, year, 2 months*, week	Yesterday, Vince remembered how much fun those times were and bought a new pair of running shoes.	Yesterday, Veronica was in a car accident and became permanently paralyzed from the waist down.
Sells companies	5 years, year		
Sells things on eBay	5 years, year		
Smokes cigarettes	year, month*, week	Yesterday, Tina wanted a smoke and bought a pack of cigarettes.	Yesterday, Vince quit smoking cigarettes.
Smokes marijuana	5 years, year*, month, week	Yesterday, William wanted to get high and bought some marijuana.	Yesterday, Vince realized he is severely allergic to marijuana and will no longer smoke it.
Steals cars	5 years*, year, month	Yesterday, William learned a new technique for breaking into cars.	Yesterday, Tina got caught and went through a radical transformation, vowing to never break the law again.
Steals chewing gum	5 years*, year, month	Yesterday, Vince learned a new trick to distract shopkeepers.	Yesterday, William got caught and vowed to never break the law again.
Volunteers for political campaigns	5 years*, year	Yesterday, Vince researched a new political candidate in the area and is going to volunteer with them.	Yesterday, Vince grew disillusioned with the political system and wants nothing to do with it anymore.
Volunteers for soup kitchens	5 years*, year	Yesterday, Vince researched a new soup kitchen in the area and is going to volunteer with them.	Yesterday, Tom grew disillusioned with the soup kitchen system and wants nothing to do with it anymore.
Watches professional football	2 years, year*, month	Yesterday, William remembered how much enjoyable that was and upgraded his cable to have access to all professional football games.	Yesterday, Veronica learned about all the corruption in professional sports and no longer can watch it.
Watches space launches	2 years*, year, month	Yesterday, William remembered how much enjoyable that was and researched all of the space launches in the next year within driving distance.	Yesterday, William went through a radical transformation and now it is against his belief to witness anything relating to space travel.
Wears a bra	6 months, month, week		
Wears a suit	6 months, month*, week	Yesterday, Vince got a high paying job on Wall Street.	Yesterday, Veronica got fired from her job on Wall Street and now works in a pizza parlor.
Wears a watch	6 months, month, week		
Wears slacks	6 months, month, week		
Wears socks	6 months, month, week		
Writes novels	5 years*, year	Yesterday, William finished an MFA program and quit his other job to focus on writing novels.	Yesterday, William became fed up with the literary world and decided to never write anything again.
Writes poems	5 years, year*	Yesterday, William finished an MFA program and quit his other job to focus on writing poems.	Yesterday, Veronica became fed up with the poetry world and decided to never write poems again.

Note. Items used in Case Study 2 (Habitual language). Referent frequency denotes the experimentally supplied time periods during which a person did an action 3 times (e.g., "In the past 5 years, John climbed mountains 3 times."; Experiment 2b). Enabling and preventative columns provide the causal manipulation sentences used in Experiment 2c to enable or prevent future instances of the action (blank entries indicate that the item was not used in Experiment 2c). Referent frequency with an asterisk denotes the time interval used in Experiment 2c.

Received June 7, 2018
Revision received October 25, 2018
Accepted October 29, 2018 ■