Supplementary Online Materials

Generic Language for Social and Animal Kinds: An Examination of the Asymmetry Between Acceptance and Inferences

Federico Cella, Kristan A. Marchak, Claudia Bianchi, Susan A. Gelman

10.1111/cogs.13209

Table of Contents

Study S1: Original Truth-Conditions Task

In addition to the Truth-Conditions and the Implied Prevalence tasks, we also conducted a direct replication of the Truth-Conditions task from Cimpian et al. (2010). In the *Expanded Truth-Conditions* (ETC) task reported in the manuscript, we included two items at the 100% prevalence level; however, this differed from the *Original Truth-Conditions* (OTC) task in Cimpian et al.'s (2010) studies. The results from the OTC without a 100% prevalence level are reported below.

Method

Participants. Fifty-nine adults from the United States (32 men, 27 women; Mean age = 39.46 years; range = 24–71 years) completed the study online via MTurk and were paid 40 cents. Participants were 76% White, 10% Asian or Asian American, 5% Black or African American, 3% Middle Eastern or North African, 2% American Indian or Alaska Native 2% Latino or Hispanic, and 2% Multiethnic. One additional participant was tested and excluded from the final sample for having a non-US IP address. Two other participants were tested and excluded for having duplicate IP addresses.

Materials and Procedure. In the OTC task, participants judged whether generic statements were "true" based on the prevalence level of an ascribed property – each of the following prevalence levels was presented twice: 10%, 30%, 50%, 70%, and 90%. Participants in the OTC task received 10 items randomly selected from the list of 12 items (see Appendix A for a complete list of the items).

Results and Discussion

Data Coding. We coded data in the OTC task in a similar way to the ETC task reported in the manuscript. However, in the OTC task, we imputed a score of 100% to participants who judged all items to be false.

We observed an asymmetry between ratings in the Implied Prevalence task and the Original Truth-Conditions task, with participants providing higher mean ratings in the Implied Prevalence Task than in the OTC task, $t(112) = 9.22$, $p < .001$, $d = 1.73$.

To examine whether participants' responses differed between the ETC and OTC tasks, we explored participants' "true/false" responses. We submitted responses to a logistic mixed-effects model using the glmer command in the lme4 package in R (Bates, 2007).[1] In this model, we included task (OTC = 0; ETC = 1; between-subject), prevalence (.1, .3, .5, .7, .9; within-subject), and their interaction as predictors (see Table S1). All predictors were mean-centered. We also included *participant* as a random intercept.[2]

Fixed Effects	**Estimate**	***SE***	***p*-value**
(Intercept)	7.65	1.83	< .001
Task	6.35	2.09	.002
Prevalence	12.24	1.24	< .001
Task x Prevalence	1.28	2.12	.55
Random Effect		**SD**	
Participant	Intercept	9.99	

Table S1. Logistic regression predicting "true" judgments, based on version of the Truth-Conditions task and prevalence of the property in Study S1.

We observed a main effect of task, with ratings higher in the ETC than in the OTC Task. We additionally observed a main effect of prevalence. However, there was no interaction

[1] In the preregistration for this study, we indicated that we would use ANOVA to analyze the data. However, since the data is binary, it is necessary to analyze the data using non-parametric statistics, such as logistic regression.

[2] We additionally fit a model including item as a random intercept; however, we found that the estimate for the SD of the intercept for item was zero, so we omitted item as a random effect in the final model.

between these factors, suggesting that participants in both versions of the task responded to items at a given prevalence level in similar ways. We thus decided to include a 100% level in subsequent tasks, because it was easier to interpret the data if we did not have to impute scores to participants who judged all items to be false.

Study 2 – Physical vs. Non-physical Properties

In Study 2, for the sake of generality, we included 12 items that described a physical property of the category (e.g., “Xs have large tonsils”), and 12 items that described a non-physical property (e.g., “Xs sleep under trees”). We thus explored whether property type (physical vs. non-physical) would affect the asymmetry and the acceptance conditions of generics.

Results and Discussion

We conducted a repeated measures analysis of variance (ANOVA) on the mean prevalence ratings, with task (Implied Prevalence vs. Truth-Conditions) and domain (animals vs. people) as between-subjects factors. We additionally included property type (physical vs. non-physical) as a within-subjects factor (see Fig. S1). We observed a main effect of task, with ratings overall higher in the *Implied Prevalence* task than the *Truth-Conditions* task, $F(1, 209) = 141.96$, $p < .001$, $\eta_p^2 = .40$. No other main effects or interactions were significant.

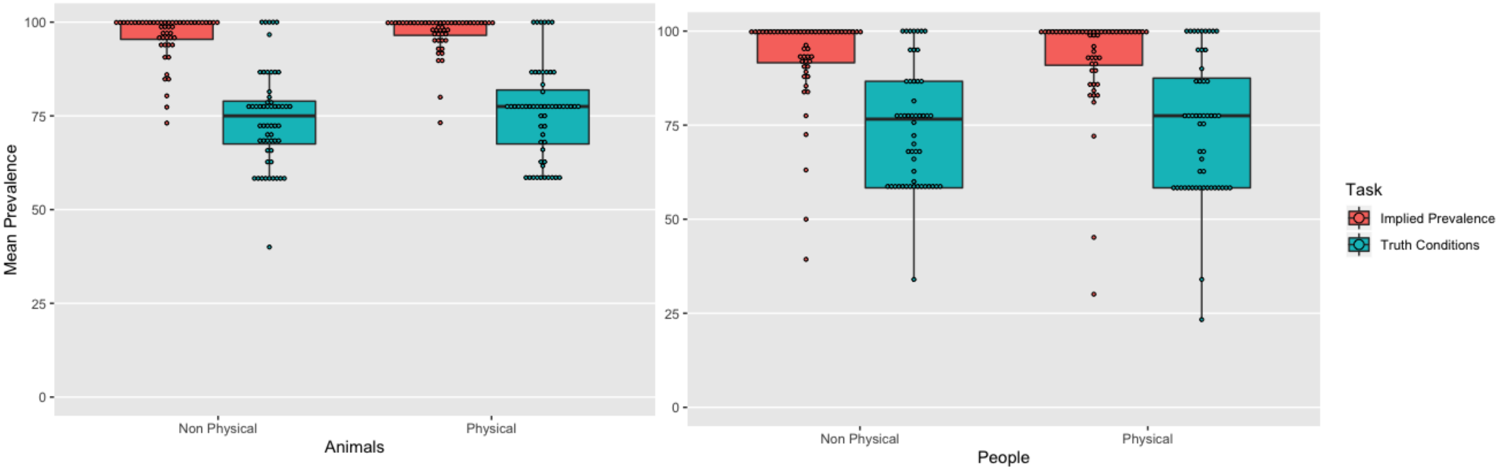

Fig. S1. Study 2, dot plots (with box plot overlays) of mean prevalence ratings, plotted by condition (Implied Prevalence vs. Truth Conditions) and property type (physical vs. non-physical). Each dot represents the mean prevalence rating for a participant, which was computed by averaging responses across the 12 physical items and 12 non-physical items that they rated. The solid line represents the median.

We also submitted participants' "true/false" responses in the Truth-Conditions task to a logistic mixed-effects model using the glmer command in the lme4 package in R (Bates, 2007). In this model, we included domain (animals = 0; people = 1; between-subject), property type (non-physical = 0; physical = 1; within-subject), prevalence (1, .3, .5, .7, .9, 1; within-subject), and their interactions as predictors (see Table S2). All predictors were mean-centered. We also included *participant* as a random intercept.[3] We observed a main effect of prevalence, indicating that generic sentences were more likely to be judged to be true for higher than lower prevalence levels. We additionally observed an interaction between domain and prevalence.

Fixed Effects	**Estimate**	***SE***	***p*-value**
(Intercept)	2.30	0.39	< .001
Domain	0.09	0.75	.90
Property Type	-0.25	0.15	.11
Prevalence	9.26	0.45	< .001
Domain x Property Type	-0.08	0.31	.79
Domain x Prevalence	-2.29	0.82	.01
Property Type x Prevalence	-0.03	0.53	.95
Domain x Property Type x Prevalence	-1.66	1.05	.12
Random Effect		**SD**	
Participant	Intercept	3.68	

Table S2. Logistic regression predicting "true"/"false" judgments, based on domain, property type, prevalence, and their interactions in Study 2.

To explore the interaction (see Fig. S2), we conducted post-hoc tests that revealed that participants were numerically, but not significantly, more likely to endorse social generics than generics about animals at lower prevalence levels: 10% level (Average Marginal Effect (AME) =

[3] We additionally fit a model including item as a random intercept; however, we found that the estimate for the SD of the intercept for item was zero, so we omitted item as a random intercept in the final model in this and subsequent studies.

0.07, SE = 0.05, *p* = .12, 95% CI = -0.02, 0.17), 30% level (AME = 0.08, SE = 0.08, *p* = .32, 95% CI = -0.07, 0.22), and 50% level, (AME = 0.03, SE = 0.08, *p* = .71, 95% CI = -0.13, 0.19). In contrast, we observed that participants were numerically, but not significantly, more likely to endorse generics about animals than social generics at higher prevalence levels: 70% level (AME = -0.01, SE = 0.05, *p* = .81, 95% CI = -0.12, 0.09), 90% level (AME = -0.03, SE = 0.04, *p* = .45, 95% CI = -0.12, 0.05), and 100% level (AME = -0.04, SE = 0.04, *p* = .32, 95% CI = -0.11, 0.04).

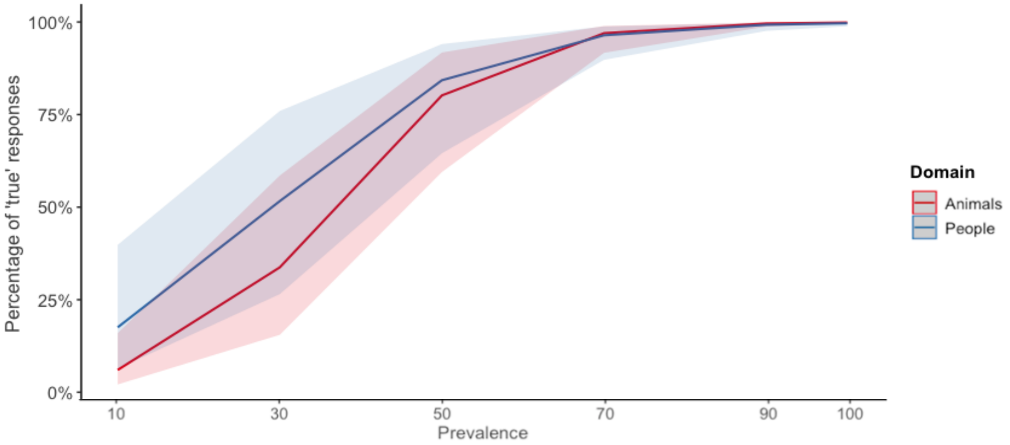

Fig. S2. Mean percentage of "true" responses in Study 2 (Truth-Conditions task) by domain and prevalence level.

These results show that the findings reported in the manuscript replicate when we include property type as a factor – property type did not affect the asymmetry nor the acceptance conditions of generics.

Study S2: Supplementary Study

We conducted a study to obtain baseline data on the homogeneity of the properties presented in Study 2, in the absence of generic language. We adapted a procedure used in previous research to observe the prevalence that participants judged our properties to have within a category (Nisbett et al., 1983).

Method

Participants. Seventy-four adults from the United States (28 men, 46 women; Mean age = 40.43 years; range = 27–67 years) completed the study online and were paid 75 cents. Participants were 78% White, 11% Black or African American, 5% Multiethnic, 4% Asian or Asian American, and 1% Latino or Hispanic. Participants were randomly assigned to either the *Animal Condition* (n = 40) or the *People Condition* (n = 34). Five participants were tested and excluded from the final sample because they failed the same manipulation check used in Study 2 (n = 4 in the *People Condition*, n = 1 in the *Animal Condition*). One participant was also tested and excluded from the final sample for not being a native speaker of English.

Materials and Procedure. Participants completed a *Prevalence Estimation* task, where they estimated the prevalence of the properties based on three members of a novel category. The structure of the items was adapted from Nisbett et al. (1983). The following is the introductory text we used for this study:

"Imagine that you are an explorer who has landed on a remote island. This island is very large, and has many different **[animals]**/**[people]** on it. It is roughly the size of Alaska, and has a lot of geographical, climatic, and environmental variety. You encounter a number of new **[animals]**/**[people]**. You observe the properties of your "samples" and you need to make guesses about how common these properties would be in other **[animals]/[people]**."

Each participant evaluated the same 24 properties used in Study 2, randomly paired with the same novel category labels. The order of the items was randomized for each participant. See Table S3 for an example of an item as it appeared to participants.

Information:
Suppose you encounter three STADES. These three STADES bury their leftover food.

Question:
What percentage of STADES on the island do you expect to bury their leftover food?

Table S3. Sample item from Study S2.

Results and Discussion

We conducted an independent sample *t*-test of the mean prevalence ratings across domains. We observed that participants provided higher mean ratings for *Animals* than *People*, *t*(63.41) = 2.46, *p* = .02, *d* = 0.58 (see Fig. S3). The results show that at baseline participants judged the animal categories in our study to be more homogeneous than the social categories (e.g., they judged members of an animal category to be more alike than members of a social category).

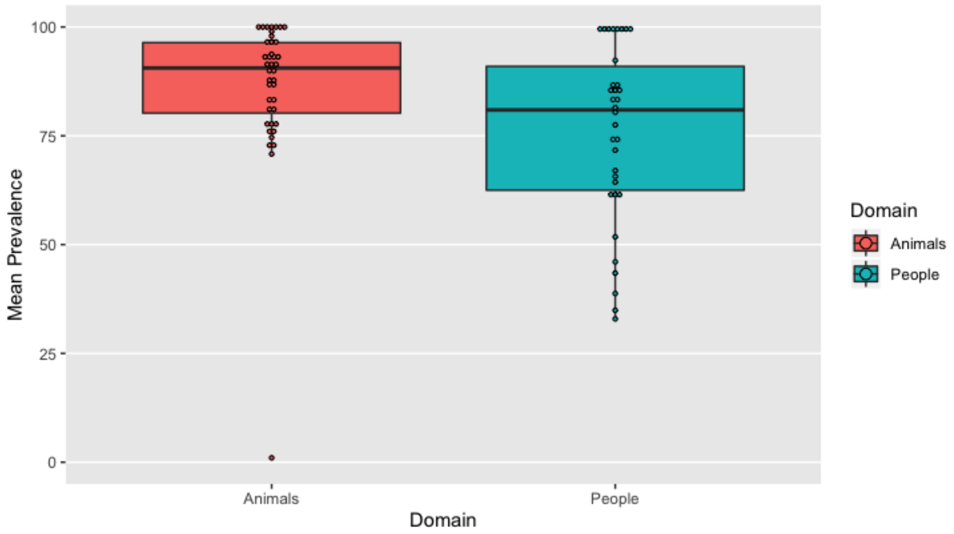

Fig. S3. Study S2, dot plots (with box plot overlays) of mean prevalence ratings, plotted by domain (Animals vs. People). Each dot represents the mean prevalence rating for a participant, which was computed by averaging responses across the 24 items. The solid line represents the median.

Study 3a: Non-Masters

We found that recruitment on MTurk slowed considerably after Study 2. To allow us to continue with data collection, we began recruiting non-Masters approximately halfway through data collection. However, when we examined the data, we observed differences between the pattern of responses across samples (Masters vs. non-Masters). To ensure that our recruitment method was consistent across studies, we replaced the non-Masters in our sample in Study 3a with Masters. Below we report the results of our primary analyses for the non-Masters sample.

Method

Participants. One-hundred and one adults from the United States (62 men, 39 women; Mean age = 35.11 years; range = 21–64 years) completed the study online and were paid $1 for participating. Participants were randomly assigned to one of four conditions: *Animals – Truth-Conditions* (n = 35), *People – Truth-Conditions* (n = 18), *Animals – Implied Prevalence* (n = 24), or *People – Implied Prevalence* (n = 24). Unlike the participants reported in the manuscript, these participants had not been granted a Master Worker qualification from MTurk. Participants were 79% White, 9% Black or African American, 5% Multiethnic, 4% Latino or Hispanic, 2% Asian or Asian American, and 1% not listed. Fifteen participants were tested and excluded from the final sample because they failed the manipulation check (n = 6 in the *People – Truth-Conditions*, n = 9 in the *People – Implied Prevalence*). Two participants were tested and excluded from the final sample for having non-US IP addresses. Finally, two participants were tested and excluded from the final sample because they were not native speakers of English.

Materials and Procedure. The materials and the procedure were the same as reported in Study 3a in the manuscript.

Results and Discussion

To explore the pattern of responses in our non-Masters sample, we conducted a repeated measures ANOVA of the mean prevalence rating, with task (Implied Prevalence vs. Truth-Conditions) and domain (animals vs. people) as between-subjects factors, and property valence (dangerous vs. neutral) as a within-subjects factor. Unlike with the Masters', we did not find a main effect of task, $F(1,97) = 2.03$, $p = .16$, $\eta_p{}^2 = .02$. However, like the Masters', we observed a main effect of property valence, $F(1,97) = 16.35$, $p < .001$, $\eta_p{}^2 = .14$, with higher ratings for neutral properties than dangerous properties and an interaction between task and property valence, $F(1, 97) = 7.31$, $p =. 008$, $\eta_p{}^2 = .07$.

Given this interaction, we examined the simple main effects of property valence within the *Implied Prevalence* and *Truth-Conditions* tasks separately. In the *Implied Prevalence* task, the mean prevalence implied by statements was higher for neutral than dangerous properties, $F(1, 97) = 22.86$, $p < .001$, $\eta_p{}^2 = .19$. In contrast, in the *Truth-Conditions* task, the mean prevalence that led participants to accept the statements did not differ between the two property valences, $F(1, 97) = 0.89$, $p = .35$, $\eta_p{}^2 = .009$.

We additionally submitted participants' responses in the *Truth-Conditions* tasks to a logistic mixed-effects model using the glmer command in the lme4 package in R (Bates, 2007). In this model, we included domain (animals = 0; people = 1; between-subject); property valence (neutral = 0; dangerous = 1; within-subject), prevalence (.1, .3, .5, .7, .9, 1; within-subject), and their interactions as predictors (see Table S3). All predictors were mean-centered. We also included *participant* as a random intercept. We observed a main effect of prevalence, indicating that generic sentences were more likely to be judged true for higher than lower prevalence levels. However, this main effect needs to be interpreted within the context of a significant property valence by prevalence interaction, which was not observed with the Masters sample (see Fig. S4).

Fixed Effects	**Estimate**	***SE***	***p*-value**
(Intercept)	2.39	0.43	< .001
Domain	0.02	0.89	.98
Property Valence	-0.36	0.28	.20
Prevalence	14.02	1.00	< .001
Domain x Property Valence	0.54	0.60	.37
Domain x Prevalence	2.51	1.93	.19
Property Valence x Prevalence	-3.74	1.16	.001
Domain x Property Valence x Prevalence	1.32	2.43	.59
Random Effect		**SD**	
Participant	Intercept	2.76	

Table S4. Logistic regression predicting "true"/"false" judgments, based on domain, property valence, prevalence, and their interactions in the non-Masters sample Study 3a.

Post-hoc tests revealed that participants were more likely to endorse generics about dangerous properties than neutral properties at the 10% level (AME = 0.07, SE = 0.02 *p* = .004, 95% CI = 0.02, 0.11) and the 30% level (AME = 0.07, SE = 0.03 *p* = .02, 95% CI = 0.01, 0.13). In contrast, there was no difference between responses to dangerous and neutral properties at higher prevalence levels: 50% level (AME = -0.009, SE = 0.03, *p* = .77, 95% CI = -0.07, 0.05), 70% level (AME = -0.05, SE = 0.02, *p* = .049, 95% CI = -0.09, -0.0001), 90% level (AME = -0.02, SE = 0.01, *p* = .09, 95% CI = -0.04, 0.003), or the 100% level (AME = -0.008, SE = 0.006, *p* = .17, 95% CI = -0.02, 0.004).

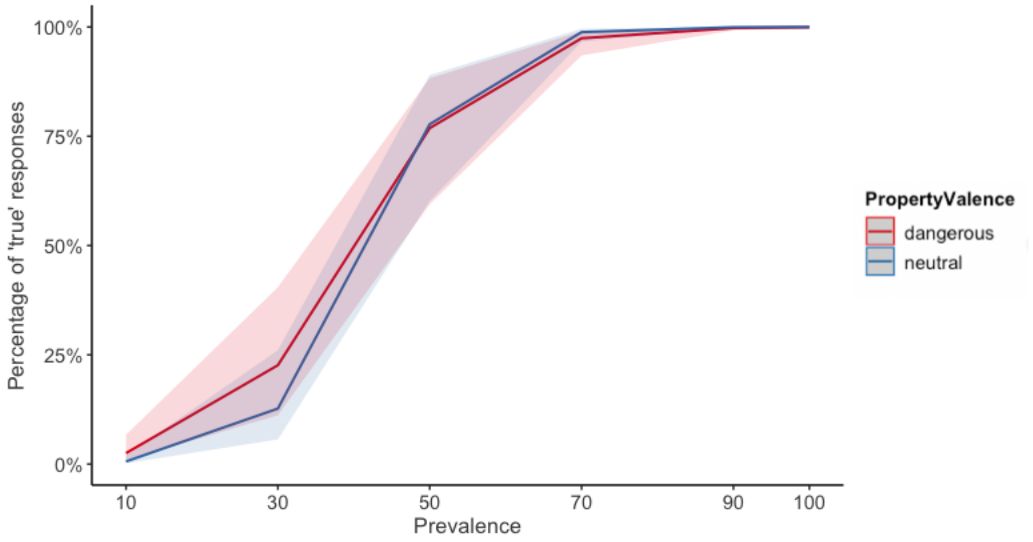

Fig. S4. Mean percentage of "true" responses in Study 3a non-Masters sample by property valence and prevalence level. For a complete overview of the mean endorsement percentages at each prevalence level by domain and property valence, see "Truth-Conditions – Study S3a" in the SOM.

Study 3a: Length Differences

Out of an abundance of caution, we conducted a supplementary analysis to ensure that the property valence effects observed in Study 3a were not due to the small differences in the length of some of our items. We controlled for predicate length by dropping the two shortest dangerous behavior items (i.e., “attack people” and “kidnap babies”) and the two longest neutral behavior items (i.e., “stamp their feet to great others” and “leave their leftovers on the ground”). The reason for this choice was that dangerous traits and neutral traits had all the same length, whereas dangerous behaviors were slightly shorter than neutral behaviors. Then, we examined participants’ responses in the *Implied Prevalence* task based on the remaining 20 items.

Results and Discussion

To explore this question, we conducted an independent sample *t*-test of the mean prevalence ratings for dangerous and neutral properties collapsed across domain. We observed a difference between dangerous and neutral properties, $t(119) = 5.94$, $p < .001$, $d = 0.54$. Even with stricter control on the length of items, judgments of implied prevalence were higher for neutral properties ($M = 92.46$, $SD = 14.85$) than dangerous properties ($M = 84.61$, $SD = 22.83$).

Study 3a: Traits vs Behavior

We were additionally interested in whether people's responses in Study 3a differed depending on whether the property described a trait or a behavior. To explore the effect of property content, we examined participants' responses in the *Implied Prevalence* task. (We could not conduct this analysis in the *Truth-Conditions* tasks, because each property was randomly assigned to a prevalence level, and thus participants did not necessarily receive an equal number of questions about traits/behaviors for both dangerous and neutral properties at the same prevalence levels.)

Results and Discussion

We conducted a repeated measures ANOVA with property valence (dangerous vs. neutral) and property content (behavior vs. trait) as within-subjects factors. We found a main effect of property valence, $F(1, 119) = 35.33$, $p < .001$, $\eta_p^2 = .23$, with higher ratings for neutral properties than dangerous properties. We additionally found a main effect of property content, with higher ratings for traits than behaviors, $F(1, 119) = 16.54$, $p < .001$, $\eta_p^2 = .12$. However, these main effects need to be interpreted within the context of a significant interaction, $F(1, 119) = 15.21$, $p < .001$, $\eta_p^2 = .11$. To further explore this interaction, we examined the simple main effects. When participants judged the implied prevalence of a dangerous property, their ratings were higher for traits than behaviors, $F(1, 119) = 20.62$, $p < .001$, $\eta_p^2 = .15$. In contrast, when participants judged a neutral property, their ratings did not differ based on the content of the property, $F(1,119) = 0.001$, $p = .98$, $\eta_p^2 = .000$.

The difference we observed may be due to the fact that traits cover a wider range of actions than behaviors. For example, a trait like "being dangerous" might be manifested in numerous ways, while a behavior like "hunting strangers" corresponds to more specific actions.

For this reason, traits might elicit higher prevalence estimates than behaviors. This finding is also compatible with the interpretation model put forward by Tessler and Goodman (2019), as naïve listeners' prior knowledge about the generalizability of traits vs. behavior might lead to higher estimates based on the former than on the latter. This result is consistent with previous evidence showing that different verbs, like "have" and "like", might have different "generalizing power" (Abelson & Kanouse, 1966; Gilson & Abelson, 1965). For example, Gilson and Abelson (1965) found that people were more likely to accept generics when they were about category members that "have sports magazines" than when they simply "like sports magazines", and this might extend to the verb "are" that we included in all traits but not in behaviors. Different verbs might thus elicit different degrees of generic asymmetries, suggesting another possible direction for future research.

We also observed that, when evaluating neutral generic properties, participants' estimates were not different for traits and behaviors. We hypothesize that this discrepancy with dangerous generic properties may be due to a ceiling effect. Indeed, participants' estimates based on neutral properties were higher than those for dangerous properties, and this might have prevented observing any difference due to property content.

Study S3a: Supplementary Study

As we did in Study S2, we explored how participants perceived the homogeneity of various category/property pairings presented in Study 3a by asking them whether the same properties could generalize from three instances of a category to other members of the same category.

Method

Participants. Seventy-seven adults from the United States (49 men, 28 women; Mean Age = 39.97 years; range = 25–67 years) completed the study online and were paid 75 cents. Unlike the participants reported in the manuscript, these participants had not been granted a Master Worker qualification from MTurk. Participants were 82% White, 10% Black or African American, 5% Asian or Asian American, 1% Latino or Hispanic, and 1% Multiethnic. Participants were randomly assigned to either the *Animal Condition* (n = 41) or the *People Condition* (n = 36). Ten participants were tested and excluded from the final sample because they failed the same manipulation check used in Study 2 (n = 8 in the *People Condition*, n = 2 in the *Animal Condition*). One additional participant was tested and excluded for having a non-US IP address. One participant was tested and excluded from the final sample for having a duplicate IP address. Finally, one participant was tested and excluded from the final sample for not being a native speaker of English.

Materials and Procedure. Participants were presented with the same introductory text as in Study S2 and completed the same *Prevalence Estimation* task. They were, however, presented with the properties tested in Study 3a.

Results and Discussion

We conducted a repeated measures ANOVA with domain (animals vs. people) as a between-subjects factor, and property valence (neutral vs. dangerous) as a within-subjects factor (see Fig. S5). The ANOVA revealed a main effect of domain, with ratings overall higher in the *Animal Condition* than the *People Condition,* $F(1,75) = 11.97$, $p = .001$, $\eta_p^2 = .14$. We also observed a main effect of property valence, $F(1,75) = 18.16$, $p < .001$, $\eta_p^2 = .20$, with higher ratings for neutral properties than dangerous properties.

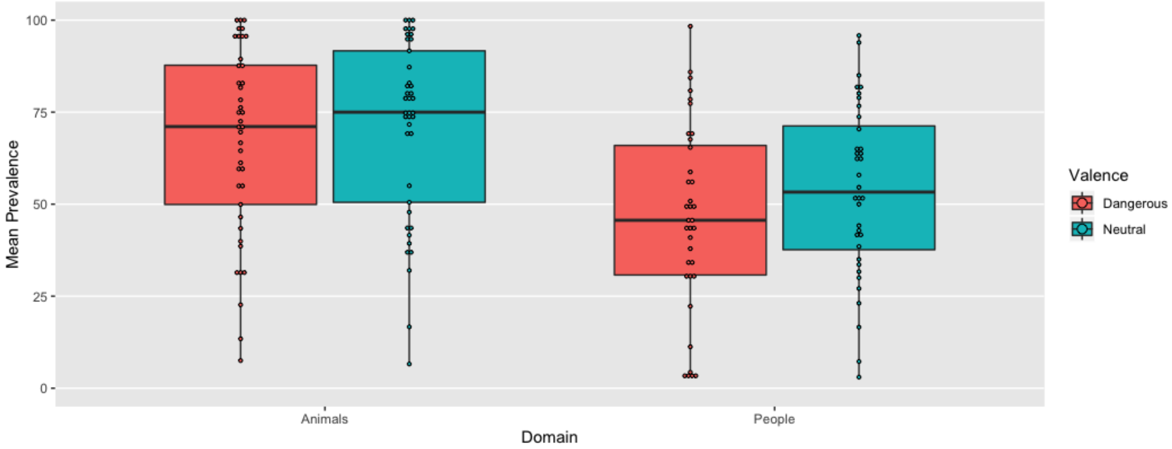

Fig. S5. Study S3a, dot plots (with box plot overlays) of mean prevalence ratings, plotted by domain (Animals vs. People). Each dot represents the mean prevalence rating for a participant, which was computed by averaging responses across the 12 dangerous items and the 12 neutral items. The solid line represents the median.

Our data thus demonstrate that animal categories are judged to be more homogenous than social categories and that neutral properties are more generalizable than dangerous properties.

Study 3b: Non-Masters

As in Study 3a, we recruited a sample of non-Masters for this study. However, when we discovered differences between the pattern of responses across samples (Masters vs. Non-Masters) in Study 3a, we also wanted to ensure that our sample was consistent across studies. Below we report the results of the non-Masters sample that we collected for this study.

Method

Participants. One-hundred and three adults from the United States (61 men, 42 women; Mean age = 35.44 years; range = 20–68 years) completed the study online and were paid $ 0.75. Participants were randomly assigned to one of two conditions: *Animals – Truth-Conditions* (n = 51) or *People – Truth-Conditions* (n = 52). Unlike the participants reported in the manuscript, these participants had not been granted a Master Worker qualification from MTurk. Participants were 77% White, 10% Black or African American, 7% Latino or Hispanic, 5% Multiethnic, and 2% Asian or Asian American. Twenty other participants were tested and excluded from the final sample because they failed the manipulation check (n = 6 in the *Animals – Truth-Conditions*, n = 14 in the *People – Truth-Conditions*). Four additional participants were tested and excluded for having non-US IP addresses. One participant was tested and excluded from the final sample for having a duplicate IP address. Finally, one participant was tested and excluded from the final sample for not being a native speaker of English.

Materials and Procedure. The materials and the procedure were the same as reported in Study 3b in the manuscript.

Results and Discussion

We submitted participants' responses to a logistic mixed-effects model.[4] In this model, we included domain (animals = 0; people = 1; between-subject); property valence (neutral = 0; dangerous = 1; within-subject), prevalence (.1, .3, .5, .7, .9, 1; within-subject), and their interactions as predictors (see Table S5). All predictors were mean-centered. We also included *participant* as a random intercept. We observed a main effect of prevalence, indicating that generic sentences were more likely to be judged true for higher than lower prevalence levels. We also observed a main effect of property valence, indicating that statements about dangerous properties were more likely to be accepted than statements about neutral properties. We did not observe this effect with the Masters sample, with whom we found a domain effect instead.

Fixed Effects	**Estimate**	***SE***	***p*-value**
(Intercept)	3.37	0.40	< .001
Domain	0.79	0.74	.29
Property Valence	0.34	0.17	.047
Prevalence	7.50	0.41	< .001
Domain x Property Valence	-0.41	0.34	.23
Domain x Prevalence	1.12	0.78	.15
Property Valence x Prevalence	-0.33	0.54	.54
Domain x Property Valence x Prevalence	-0.66	1.08	.54
Random Effect		**SD**	
Participant	Intercept	3.38	

Table S5. Logistic regression predicting "true"/"false" judgments, based on domain, property valence, prevalence, and their interactions in non-Master workers in Study S3b.

For a complete overview of the mean endorsement percentages at each prevalence level by domain and property valence, see "Truth-Conditions – Study S3b" in the SOM.

[4] In the pre-registration for this study, we indicated that we would use ANOVAs to analyze the data. However, because the data are binary, it was necessary to analyze the data using non-parametric statistics; we thus opted for logistic regression.

Study S3b: Supplementary Study

As we did in Study S2 and Study S3a, we explored how participants perceived the homogeneity of various category/property pairings presented in Study 3b.

Method

Participants. Seventy-seven adults from the United States (40 men, 37 women; M = 35.92 years; range = 23-70 years) completed the study online and were paid 75 cents. Unlike the participants reported in the manuscript, these participants had not been granted a Master Worker qualification from MTurk. Participants were 78% White, 9% Black or African American, 6% Multiethnic, 5 % Latino or Hispanic, and 1% Asian or Asian American. Participants were randomly assigned to either the *Animal Condition* (n = 40) or the *People Condition* (n = 37). Twelve participants were tested and excluded from the final sample because they failed the same manipulation check used in Study 2 (n = 6 in the *People Condition*, n = 6 in the *Animal Condition*). One additional participant was tested and excluded from the final sample for not being a native speaker of English.

Materials and Procedure. Participants were presented with the same introductory text as in Study S2 and completed the same *Prevalence Estimation* task. They were, however, presented with the properties tested in Study 3b.

Results

We conducted a repeated measures ANOVA with domain (animals vs. people) as a between-subjects factor, and property valence (neutral vs. dangerous) as a within-subjects factor (see Fig. S6). We found a main effect of domain, with ratings overall higher in the *Animal Condition* than the *People Condition,* $F(1,75) = 7.42$, $p = .008$, $\eta_p{}^2 = .09$. We also observed a main effect of property valence, $F(1,75) = 8.91$, $p = .004$, $\eta_p{}^2 = .11$, with higher ratings for neutral properties than dangerous properties. However, these two main effects needed to be interpreted within the context of a significant interaction between domain and property valence, $F(1, 75) = 11.78$, $p = .001$, $\eta_p^2 = .14$. Given this interaction, we examined the simple main effects of property valence in the *Animal* Condition and in the *People* Condition separately. In the Animal Condition, the mean prevalence estimate did not differ between neutral and dangerous properties, $F(1, 75) = 0.12$, $p = .75$, $\eta_p^2 = .00$. In the People Condition, instead, the mean prevalence estimate was higher for neutral properties than dangerous properties, $F(1, 75) = 19.81$, $p < .001$, $\eta_p^2 = .21$.

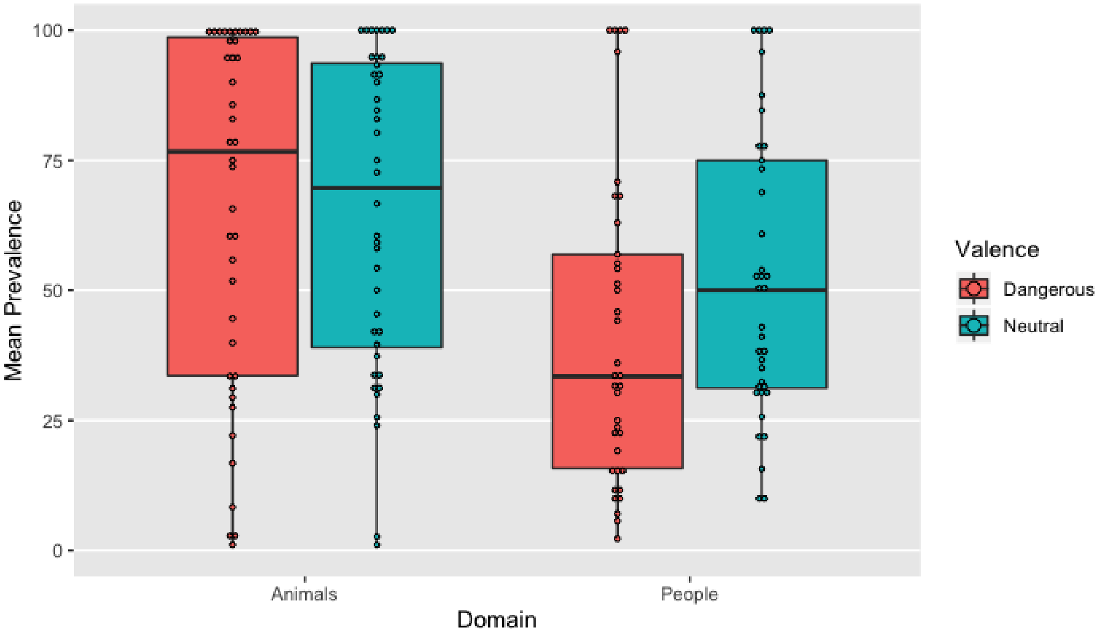

Fig. S6. Study S3b, dot plots (with box plot overlays) of mean prevalence ratings, plotted by domain (Animals vs. People). Each dot represents the mean prevalence rating for a participant, which was computed by averaging responses across the 12 dangerous items and the 12 neutral items. The solid line represents the median.

LMER Syntax for Logistic Regression Analyses

All models were calculated using lme4 package version 1.1-21 in R.

Study 2 regression (in lmer syntax)
Response ~ Domain *Prevalence + (1|Participant)

Study 3a regression (in lmer syntax)
Response ~ Domain *Prevalence*Property Valence + (1|Participant)

Study 3b regression (in lmer syntax)
Response ~ Domain *Prevalence*Property Valence + (1|Participant)

Study S1 - Mean Endorsement Truth-Conditions

		Study 1	
		OTC task	ETC task
Prevalence	10%	35.3%	44.0%
	30%	52.6%	56.0%
	50%	75.0%	79.3%
	70%	81.0%	82.8%
	90%	82.8%	81.9%

Table S6. Mean percentage of “true” responses in the OTC and ETC tasks of Study S1 at the 10–90% prevalence levels.

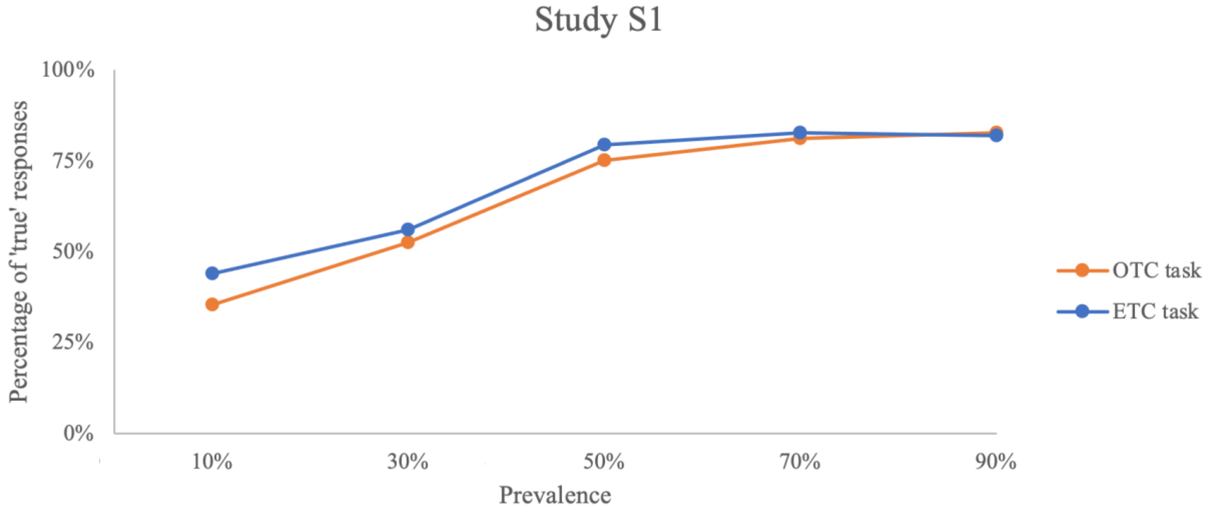

Fig. S7. Mean percentage of “true” responses in the OTC and ETC tasks of Study S1 at the 10–90% prevalence levels.

Study 2 - Mean Endorsement Truth-Conditions

Study 2					
		Animals		People	
		Physical	Non-physical	Physical	Non-physical
Prevalence	10%	20.5%	22.3%	38.5%	35.6%
	30%	31.3%	40.2%	42.3%	44.2%
	50%	68.8%	69.6%	61.5%	63.5%
	70%	85.7%	84.8%	72.1%	77.9%
	90%	84.8%	86.6%	79.8%	82.7%
	100%	98.2%	97.3%	96.2%	98.1%

Table S7. Mean percentage of "true" responses in Study 2 at each prevalence level by domain and property type.

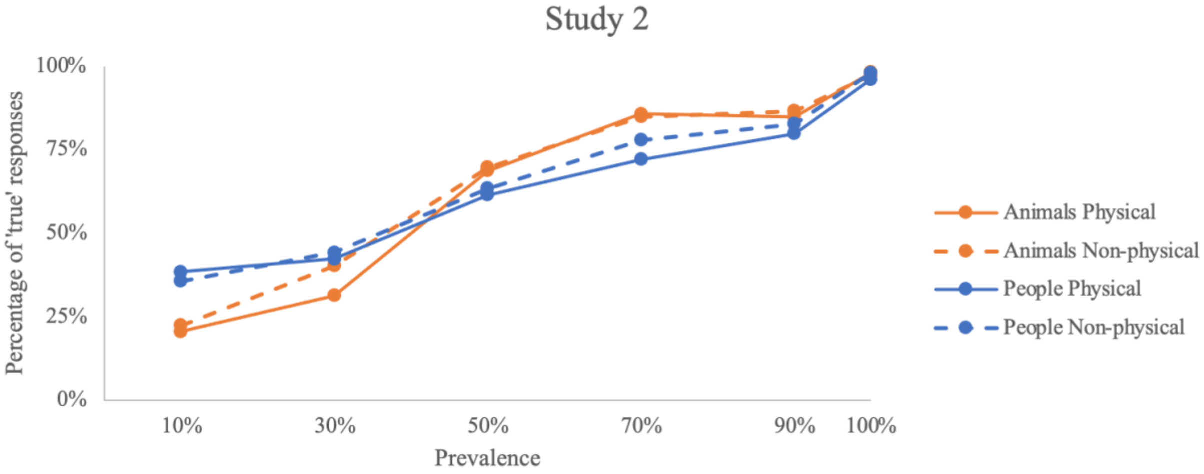

Fig. S8. Mean percentage of "true" responses in Study 2 at each prevalence level by domain and property type.

Study 3a - Mean Endorsement Truth-Conditions

Study 3a					
		Animals		People	
		Neutral	Dangerous	Neutral	Dangerous
Prevalence	10%	15.8%	21.1%	29.8%	24.0%
	30%	30.7%	32.5%	32.7%	34.6%
	50%	66.7%	71.9%	72.1%	68.3%
	70%	90.4%	92.1%	89.4%	86.5%
	90%	92.1%	93.0%	88.5%	87.5%
	100%	100.0%	100.0%	94.2%	93.3%

Table S8. Mean percentage of "true" responses in Study 3a at each prevalence level by domain and property valence.

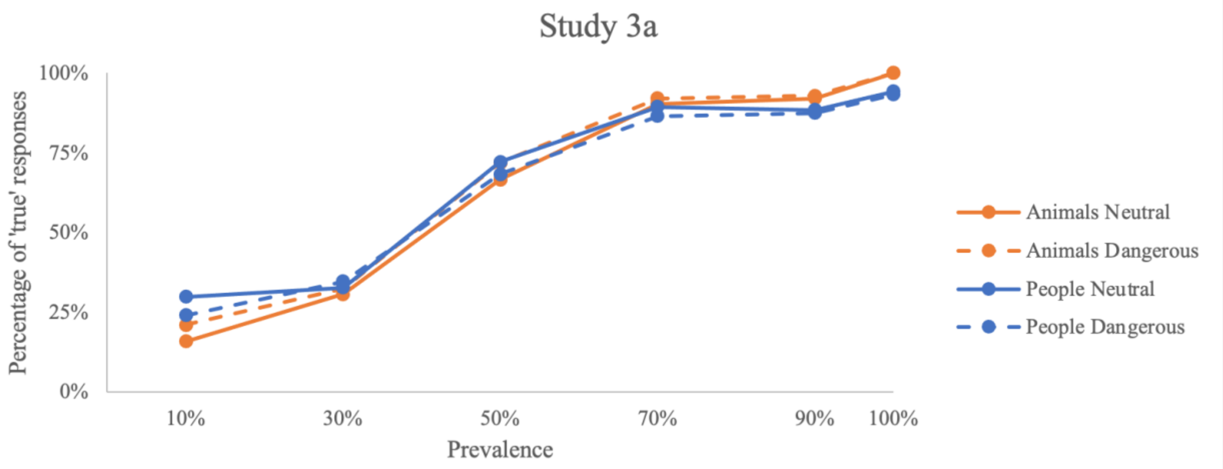

Fig. S9. Mean percentage of "true" responses in Study 3a at each prevalence level by domain and property valence.

Study 3a Mean Endorsement Truth-Conditions Non-Masters

		Study S3a (non-Masters)			
		Animals		People	
		Neutral	Dangerous	Neutral	Dangerous
Prevalence	10%	10.0%	18.6%	5.6%	13.9%
	30%	21.4%	25.7%	27.8%	27.8%
	50%	70.0%	65.7%	52.8%	61.1%
	70%	94.3%	94.3%	94.4%	91.7%
	90%	98.6%	97.1%	100.0%	94.4%
	100%	100.0%	95.7%	97.2%	100.0%

Table S9. Mean percentage of "true" responses in Study S3a at each prevalence level by domain and property valence for the non-Masters sample.

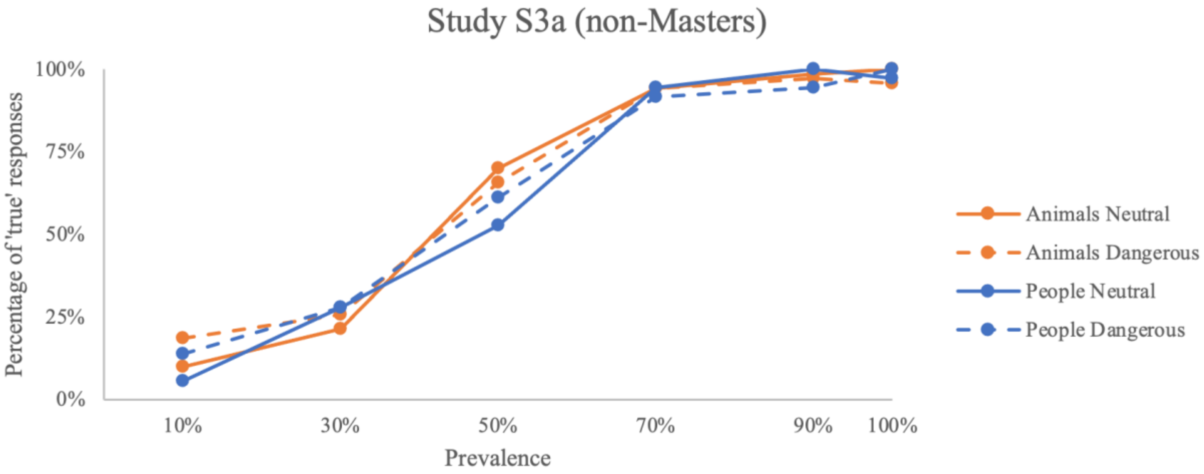

Fig. S10. Mean percentage of "true" responses in Study S3a at each prevalence level by domain and property valence for the non-Masters sample.

Study 3b Mean Endorsement Truth-Conditions

		Study 3b			
		Animals		People	
		Neutral	Dangerous	Neutral	Dangerous
Prevalence	10%	35.3%	44.0%	55.7%	59.4%
	30%	52.6%	56.0%	67.0%	71.7%
	50%	75.0%	79.3%	87.7%	86.8%
	70%	81.0%	82.8%	89.6%	91.5%
	90%	82.8%	81.9%	94.3%	94.3%
	100%	100.0%	97.4%	99.1%	100.0%

Table S10. Mean percentage of "true" responses in Study 3b at each prevalence level by domain and property valence.

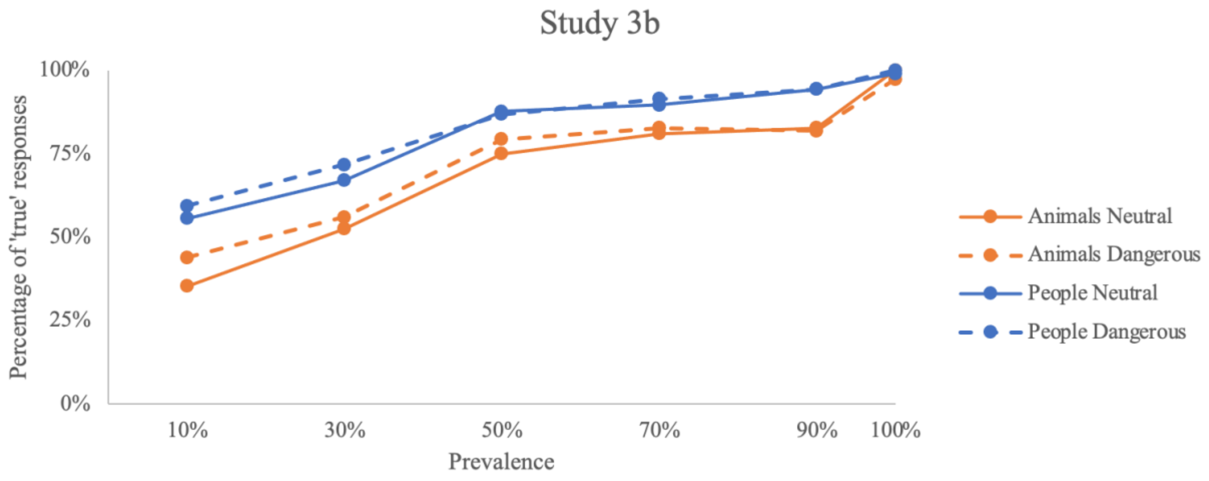

Fig. S11. Mean percentage of "true" responses in Study 3b at each prevalence level by domain and property valence.

Study 3b Mean Endorsement Truth-Conditions Non-Masters

		Study S3b (non-Masters)			
		Animals		People	
		Neutral	Dangerous	Neutral	Dangerous
Prevalence	10%	35.3%	42.2%	48.1%	51.9%
	30%	50.0%	57.8%	60.6%	64.4%
	50%	78.4%	81.4%	76.0%	76.0%
	70%	88.2%	93.1%	85.6%	85.6%
	90%	93.1%	95.1%	86.5%	87.5%
	100%	93.1%	95.1%	100.0%	99.0%

Table S11. Mean percentage of "true" responses in Study S3b at each prevalence level by domain and property valence for the non-Masters sample.

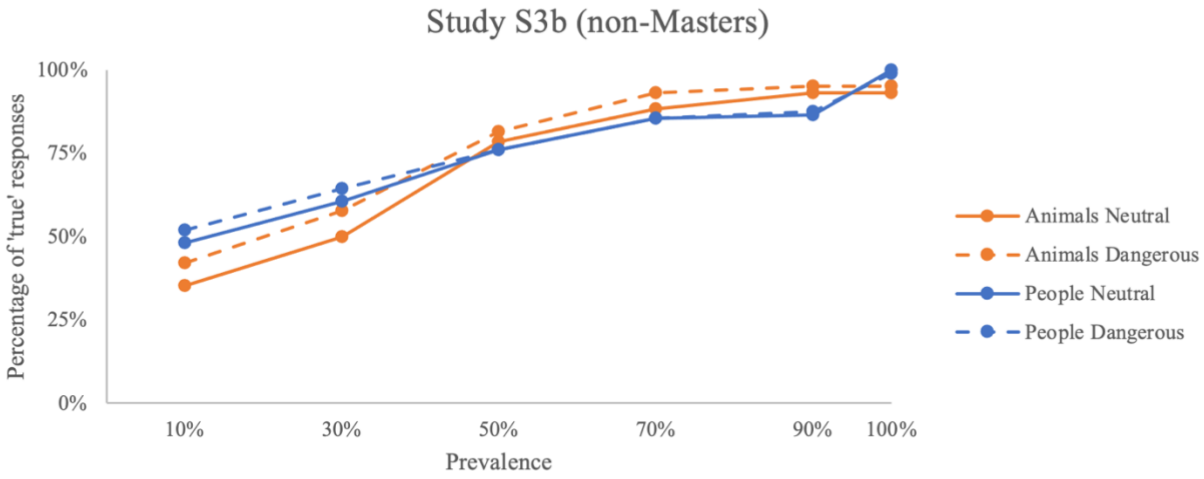

Fig. S12. Mean percentage of "true" responses in Study S3b at each prevalence level by domain and property valence for the non-Masters sample.

References

Abelson, R. P., & Kanouse, D. E. (1966). Subjective acceptance of verbal generalizations. In S. Feldman (Ed.), *Cognitive consistency: Motivational antecedents and behavioral consequents* (pp. 171–197). New York: Academic Press.

Bates, D. (2007). *Linear mixed model implementation in lme4*. University of Wisconsin–Madison

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

Gilson, C., & Abelson, R. P. (1965). The subjective use of inductive evidence. *Journal of Personality and Social Psychology*, *2*(3), 301–310.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, 90(4), 339–363.

Tessler, M., & Goodman, N. (2019). The language of generalization. *Psychological Review*, *126*(3), 395–436.