

Supplemental Materials: Using contrastive inferences to learn about new words and categories

Claire Augusta Bergey and Daniel Yurovsky

S1 Experiment 1

S1.1 Unique target display trials

In Experiment 1, half of the trials were filler trials in which the target had both a unique shape and unique value of the critical feature. We expected participants to choose the unique object when directed to “Find the toma” or “Find the blue toma” in these trials. These were included as filler trials to keep participants from cluing into the intended inference in contrastive display trials, and they provide a sanity check that people are making sensible selections in the task.

To analyze these trials, we asked whether participants chose the target more often than expected by chance (33%) by fitting a mixed effects logistic regression with an intercept term, a random effect of subject, and an offset of $\text{logit}(1/3)$ to set chance probability to the correct level. The intercept term was reliably different from zero for both color ($\beta = 6.64, t = 4.1, p < .001$) and size ($\beta = 2.25, t = 6.91, p < .001$), indicating that participants consistently chose the unique object on the screen when given an instruction like “Find the (blue) toma” or “Find the (big) toma,” across utterances with and without an adjective.

To test whether the utterance type (with or without an adjective) and feature type (size or color) affected people’s referent choices, we fit a mixed effects logistic regression predicting target selection from feature type, utterance type, and their interaction with random effects of participants. Use of an adjective in the utterance increased target choice ($\beta = 3.85, t = 3.52, p < .001$), and feature type (color vs. size) was not statistically related to target choice ($\beta = -0.48, t = -1.1, p = .269$). The two effects had a marginal interaction ($\beta = -2.24, t = -1.95, p = .051$). Participants had a general tendency to choose the target in unique target display trials, which was strengthened if the audio instruction contained the relevant adjective. These effects did not significantly differ between color and size adjectives, which suggests that participants did not treat color and size differently in these baseline trials, though the marginal interaction suggests that use of an adjective may strengthen their tendency to choose the unique object more powerfully in the size condition.

S1.2 Additional analyses

In addition to the analyses reported in the main text, we ran a pre-registered linear mixed effects model predicting target choice from the presence of an adjective in the utterance, the adjective type (size or color), and the display type (unique target display or contrastive display) (Table S1). People were more likely to choose the target if there was an adjective in the utterance ($\beta_{\text{adjective}} = 2.21, t = 7.18, p < .001$), and were more overall likely to choose the target on unique target trials ($\beta_{\text{unique}} = 3.81, t = 10.68, p < .001$). There was an interaction between the presence of

an adjective and the type of adjective, such that people were especially likely to choose the target when there was a size adjective in the utterance ($\beta_{\text{adjective*size}} = 0.95$, $t = 2.18$, $p = .029$). There was a three-way interaction between the presence of an adjective, the type of adjective, and the search type such that the contrastive strength of size over color was stronger in the contrastive trials than the unique target trials ($\beta_{\text{adjective*size*unique}} = -3.06$, $t = -2.61$, $p = .009$).

Table S1: Full model of target choice from Experiment 1. Model specification is $\text{chose_target} \sim \text{utterance_type} * \text{adjective_type} * \text{display_type} + (1 + \text{utterance_type} | \text{subject})$.

term	estimate	z-value	p-value
intercept	-2.07	-7.94	< .001
utterance type: adjective (vs. no adjective)	2.21	7.18	< .001
adjective type: size (vs. color)	-0.17	-0.46	.646
display type: unique target (vs. contrastive)	3.81	10.68	< .001
adjective * size	0.95	2.18	.029
adjective * unique target	1.32	1.22	.223
size * unique target	-0.17	-0.37	.709
adjective * size * unique target	-3.06	-2.61	.009

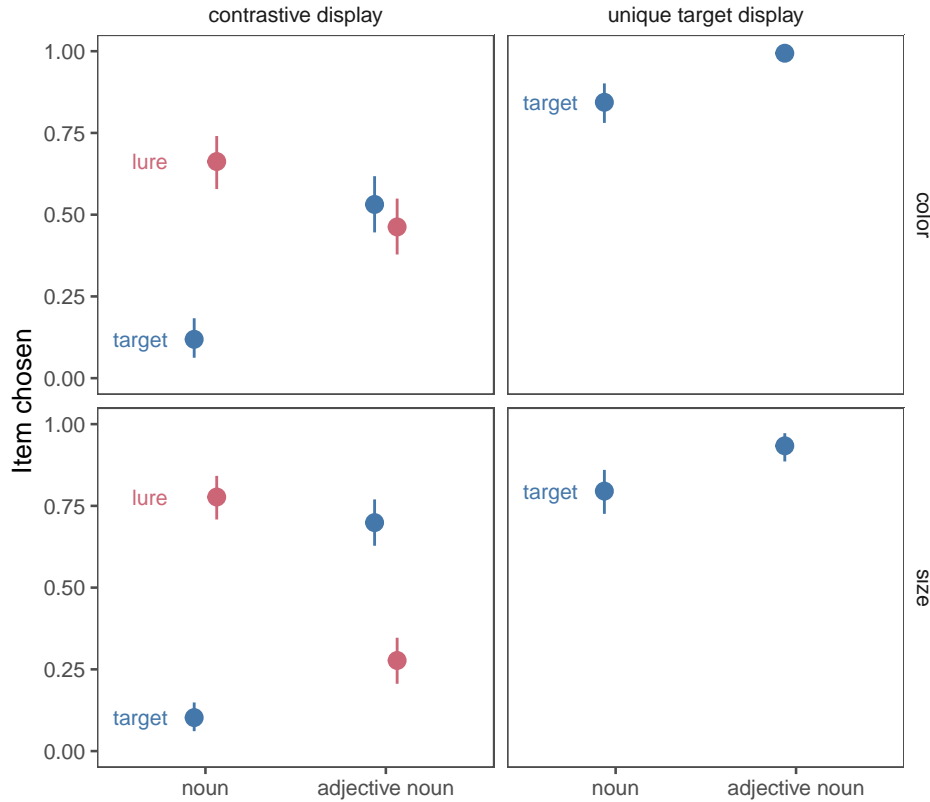


Figure S1: Referent choice in both the contrastive display trials and the unique target display trials.

Figure S1 shows referent choice in both the unique target display trials and the contrastive display trials. Unique target displays had one unique referent (the target) and two identical distractors that differed from it both in shape and the critical feature. Contrastive displays had a target, a contrastive pair which matched the target in shape but had a different critical feature, and a lure which matched the target on the critical shape but differed from it on the critical feature.

S1.3 Modeling Experiment 1 with continuous semantics

Degen, Hawkins, Graf, Kreiss, & Goodman (2020) capture asymmetries in description of size and color by positing that different features have different semantic strength. They posit that color has stronger semantics than size, such that “red table” is a better literal description of a small red table than “small table” is. Under these assumptions, RSA using these continuous semantics explains people’s tendency to mention color more often than size in a variety of tasks. Can continuous semantics explain the asymmetry we find between color and size in Experiment 1?

In Experiment 1, we found that people more consistently choose the target using contrastive inferences about size than color. We incorporated their continuous semantics into our RSA model of referent choice, which reasons over possible lexicons. In Figure S2 we show the difference in referent choice when a feature has low semantic strength (0.8) compared to high semantic strength (0.99). A feature with low semantic strength results in a weaker contrastive inference (reduced choice of the target in the *adjective noun* trials) compared to a feature with high semantic strength.

Degen et al. (2020) find that color has stronger semantics than size, which would result in a stronger contrastive inference about referent choice when color adjectives are used. This is not what we find: people make stronger contrastive inferences about referent choice when *size* adjectives are used. Thus, while a model with continuous semantics could in principle explain the asymmetry we find, it would need to have stronger semantic values for size than color. Thus, continuous semantics, when integrated into our model alone, are not a good explanation of the effect we find. In the next section, we explain how differences in production expectations—which could be generated by a continuous semantics speaker model—could explain the asymmetry we find.

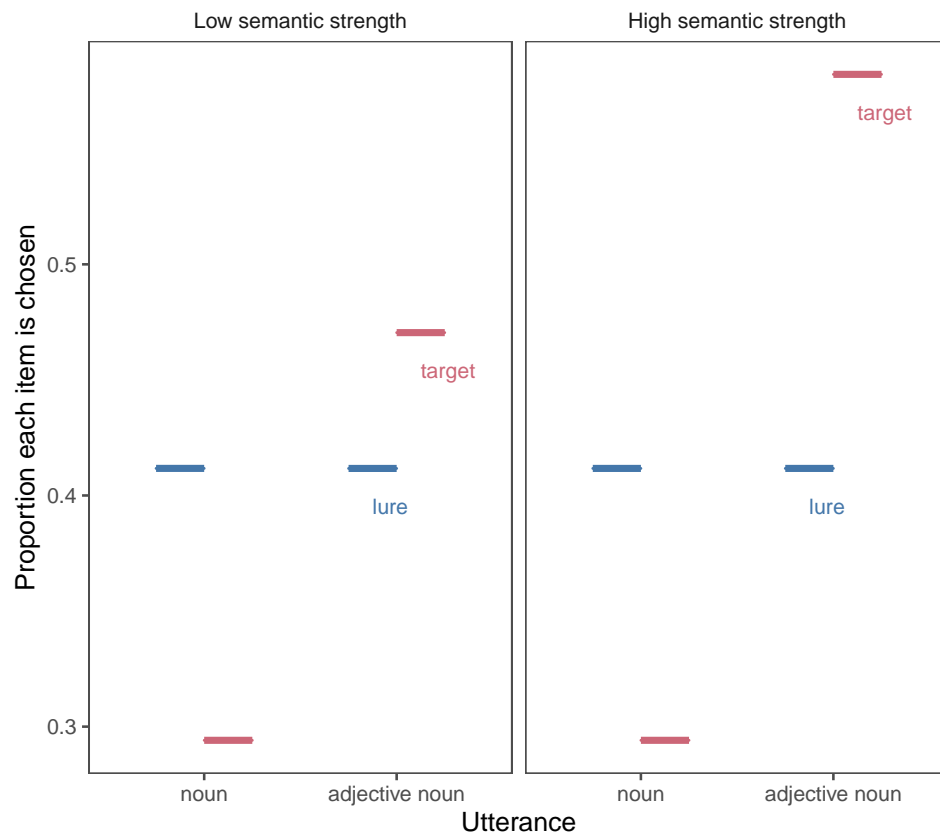


Figure S2: Results of modeling target choice in Experiment 1 using continuous semantics. Stronger continuous semantics predict higher choice of the target, while weaker continuous semantics predict lower choice of the target.

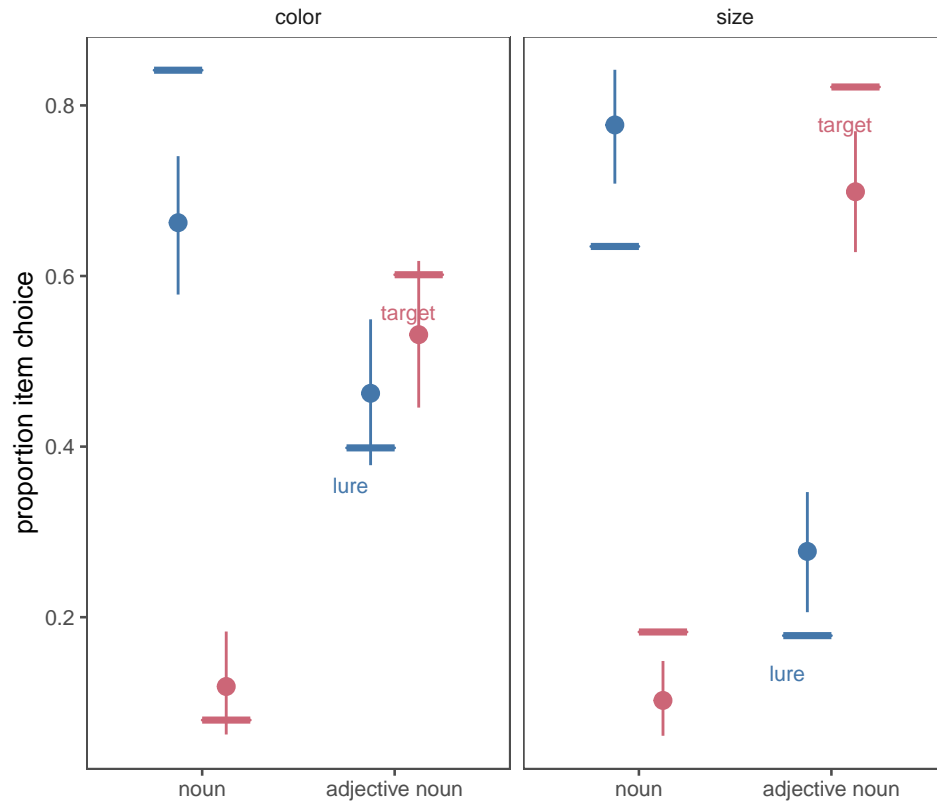


Figure S3: Results of modeling Experiment 1 using differing production probabilities. When an adjective is likely to be produced (left), the model predicts a weaker preference for the target when an adjective is produced (*adjective noun* utterance) than it does when an adjective is unlikely to be produced.

S2 Modeling Experiment 1 with Production Expectations

One possible explanation of the asymmetry we see between people’s inferences about color and size is that listeners expect speakers to mention objects’ colors more than their sizes. Speakers do remark on color more than size (Pechmann, 1989), and listeners may keep track of these expectations when interpreting utterances, downweighting the communicative weight of color adjectives. In our models in the main text, we implement this downweighting in a feature-specific rationality parameter: we posit that listeners expect speakers to be less strictly informative when talking about color than size. Another way to implement this difference is to put it in a different part of the model: the utterance prior. That is, to build into the model the listener’s expectation that speakers will mention color adjectives more often than size adjectives.

Here, we provide a demonstration that this implementation would explain the asymmetry we observe in people’s data. Figure S3 shows the model predictions when using an adjective is likely (on the left) and using an adjective is unlikely (on the right). When the listener thinks a speaker is fairly likely to use a type of adjective (as they would with color adjectives) and the adjective is used, they prefer the target, but only weakly. When a listener thinks a speaker is unlikely to use a type of adjective (as they would with size adjectives) and the adjective is used, they prefer the target very strongly. We also plot the data from people’s judgments here for comparison, but note that this demonstration is only meant to show qualitative patterns—we did not estimate specific parameters (the rationality parameter or specific rates of adjective use) to fit these data more closely.

Where might these production differences come from? There are several potential explanations of why people remark on color more than size. One explanation continuous semantics (Degen et al., 2020). That is, if color has stronger semantics than size, people should remark on color more often than size. In sum, though implementing continuous semantics directly in the literal listener of our model does not explain people’s color–size asymmetry in this task, using continuous semantics to generate production probabilities and then integrating those into the pragmatic speaker in our model can explain the asymmetry. The models we present in the main text implement this asymmetry in a feature rationality parameter for simplicity, but this alternative modeling choice is a promising avenue for explaining differences between inferences about different features in a more principled way.

S3 Experiment 2 Additional Analyses

S3.1 Experiment 2 Prevalence Judgments

The full regression of prevalence judgments, also reported in the main text, is in Table S2.

Table S2: Full model of prevalence judgments from Experiment 2. Model specification is percentage \sim adjective_type * utterance_type * context_type + (utterance_type | subject).

term	estimate	z-value	p-value
intercept	52.16	22.40	< .001
adjective type: size (vs. color)	4.73	1.46	.146
utterance type: adjective (vs. no adjective)	-10.22	-3.37	.001
context: within-category contrast display (vs. between-category contrast)	3.92	1.63	.104
context: same feature display (vs. between-category contrast)	-1.48	-0.62	.537
size * adjective	0.04	0.01	.993
size * within-category contrast display	-1.37	-0.41	.684
size * same feature display	-0.60	-0.18	.859
adjective * within-category contrast display	-1.58	-0.46	.644
adjective * same feature display	2.13	0.63	.532
size * adjective * within-category contrast display	-1.39	-0.29	.770
size * adjective * same feature display	-1.59	-0.33	.739

S4 Experiment 3 Additional Analyses

The full regression predicting Experiment 3 prevalence judgments, also reported in the main text, is shown in Table S3. The regression predicting Experiment 3 prevalence judgments among only adjective utterances and no adjective utterances (excluding alien utterance trials), also reported in the main text, is shown in Table S4.

In addition to the regressions reported in the manuscript, we report two pre-registered, targeted regressions to test the effect of utterance type to more specifically in case these effects were unclear in the maximal models. First, we filtered to adjective and no adjective trials and fit a linear mixed effects model predicting prevalence judgment by utterance type with a random slope of utterance type by subject (Table S5). Participants' prevalence judgments were significantly lower when an adjective was used in the utterance ($\beta = -9.17$, $t = -7.09$, $p = < .001$). Second, we included all trials in a linear mixed effects model predicting prevalence judgment by utterance type with a random slope of utterance type by subject (Table S6). Utterances without an adjective resulted in significantly higher prevalence judgments than alien utterances ($\beta = 7.76$, $t = 4.91$, $p = < .001$), and utterances with an adjective did not result in significantly different prevalence judgments than alien utterances ($\beta = -1.42$, $t = -0.91$, $p = .363$).

Table S3: Regression predicting prevalence judgments from utterance type, context type, and adjective type in Experiment 3. Model specification is percentage \sim utterance_type * context_type * adjective_type + (utterance_type | subject).

term	estimate	z-value	p-value
intercept	47.77	20.73	< .001
utterance type: no adjective utterance (vs. alien utterance)	7.48	2.80	.005
utterance type: adjective (vs. alien utterance)	-0.64	-0.24	.808
context type: within-category contrast (vs. between-category)	-2.70	-1.23	.220
adjective type: size (vs. color)	4.44	1.33	.185
no adjective utterance * within-category contrast display	5.57	1.79	.073
adjective utterance * within-category contrast display	5.77	1.86	.064
no adjective utterance * size	-5.09	-1.32	.189
adjective utterance * size	-6.56	-1.72	.086
within-category contrast display	1.24	0.39	.696
no adjective utterance * within-category contrast display * size	-0.32	-0.07	.944
	-2.21	-0.49	.623

Table S4: Regression predicting prevalence judgments from utterance type, context type, and adjective type only among adjective and no adjective utterances (excluding alien utterances) in Experiment 3. Model specification is percentage \sim utterance_type * context_type * adjective_type + (utterance_type | subject).

term	estimate	z-value	p-value
intercept	55.24	23.76	< .001
utterance type: adjective (vs. no adjective)	-8.12	-3.46	.001
context type: within-category contrast (vs. between-category)	2.87	1.34	.180
adjective type: size (vs. color)	-0.66	-0.20	.845
adjective utterance * within-category contrast display	0.19	0.06	.949
adjective utterance * size	-1.47	-0.43	.665
within-category contrast display	0.92	0.30	.766
no adjective utterance * within-category contrast display * size	-1.90	-0.43	.665

Table S5: Regression predicting prevalence judgments from presence of an adjective in the utterance (excluding alien language utterances) in Experiment 3. Model specification is percentage \sim utterance_type + (utterance_type | subject).

term	estimate	z-value	p-value
intercept	56.59	38.00	< .001
adjective utterance (vs. no adjective utterance)	-9.17	-7.09	< .001

Table S6: Regression predicting prevalence judgments from utterance type in Experiment 3. Model specification is percentage \sim utterance_type + (utterance_type | subject).

term	estimate	z-value	p-value
intercept	48.83	33.18	< .001
no adjective utterance (vs. alien utterance)	7.76	4.91	< .001
adjective utterance (vs. alien utterance)	-1.42	-0.91	.363

S5 Combined Analyses of Experiment 2 and 3

Given that many participants failed the memory check in Experiments 2 and 3 (which was our pre-registered exclusion criterion), we yielded data from fewer participants than expected in our pre-registrations. This means that the analyses of Experiments 2 and 3 may be underpowered to detect subtle effects of context or feature type. Since Experiments 2 and 3 had very similar procedures, we can conduct an exploratory combined analysis comparing the conditions they shared.

Combining data from Experiment 2 ($N = 193$) and Experiment 3 ($N = 197$) yields a total sample of 390 participants in this combined analysis. Only trials from the conditions that Experiments 2 and 3 share are included in this analysis: two types of referential context (within-category context and between-category context), two types of feature (size and color), and two types of utterance (adjective noun and noun only).

We fit a mixed effects linear regression on the combined data with the same specification as the full models we used for Experiments 2 and 3 in the main text: effects of utterance type, context type, and critical feature and their interactions, and a random slope of utterance type by subject. This model revealed a significant effect of utterance type ($\beta_{adjective} = -9.04$, $t = -4.68$, $p < .001$): people inferred that a feature was rarer when it was mentioned, consistent with our findings in each experiment separately. Participants' inferences did not significantly differ between color and size adjective conditions ($\beta_{size} = 2.01$, $t = 0.85$, $p = .396$). This is also consistent with findings from each experiment. There was a significant effect of context, with people making overall slightly higher prevalence judgments in the within-category context ($\beta_{within} = 3.39$, $t = 2.1$, $p = .036$), which was not found in either experiment separately. However, there was not a significant interaction between context and utterance type ($\beta_{within*adjective} = -0.74$, $t = -0.32$, $p = .748$), consistent with findings from each experiment. That is, people did not modulate the size of their typicality inferences (the difference between the adjective noun and noun utterance types) based on context type, though there was a baseline difference in prevalence judgments between context types.

Overall, these results are consistent with what we found in Experiments 2 and 3, with the exception of finding an overall difference in prevalence judgments between context types. However, we did not find an interaction between context type and utterance type, which would demonstrate that participants are trading off between referential utility and typicality when making these inferences. This exploratory combined analysis suggests further research is necessary, as even when combined our data are not definitive about whether potential trade-offs are small or nonexistent.

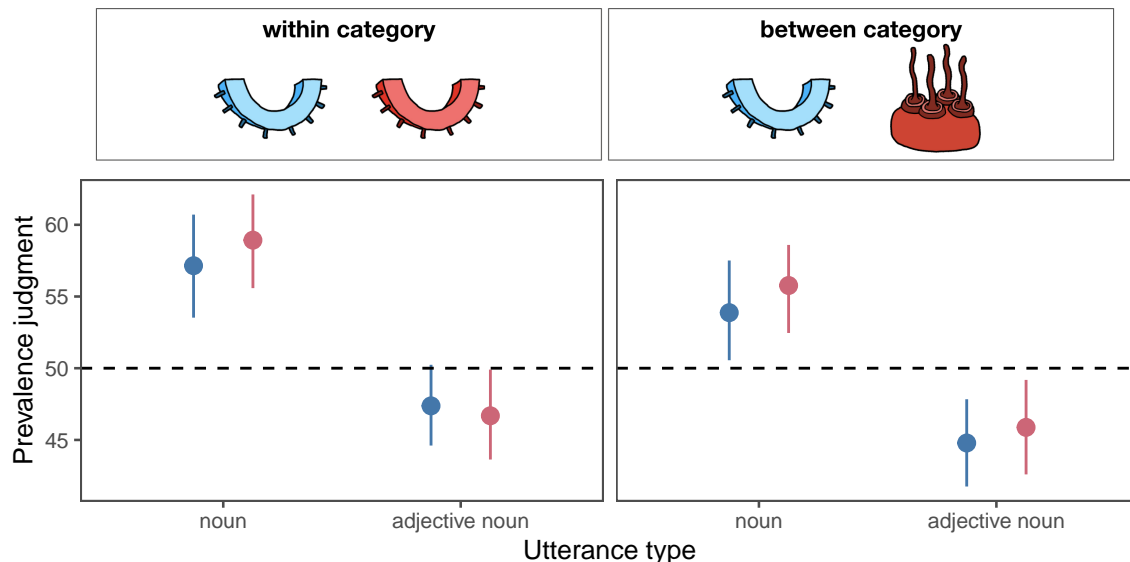


Figure S4: Participants’ prevalence judgments in combined data from Experiments 2 and 3.

S6 Inferring Atypicality with Non-Fruit Novel Stimuli

The use of alien fruit stimuli in our experiments raises the question of whether people would infer that *non-fruits* are atypical when their color is remarked upon. As noted in the introduction, people remark on the colors of some kinds of objects more than others, perhaps because they more often vary in color (Rubio-Fernández, 2016). Because fruits tend to have stereotypical colors, it is possible that people would not expect the colors of fruits to be remarked upon very often, and thus have a stronger inference of atypicality for fruit than other types of objects. Here, we provide some additional data showing that people still make these inferences for categories about which they likely do not have such an expectation: block shapes.

S6.0.1 Participants.

346 participants were recruited on Amazon Mechanical Turk to perform this task. They were paid 10 cents to complete the task. Participants on average took 20 seconds to complete the single trial (not including reading the consent form).

Two participants were excluded because they repeated the task, leaving 344 participants for analysis. There were no other exclusion criteria.

S6.0.2 Procedure

Before the main task began, participants were shown an array of colorful block shapes and told they were going to learn about shapes like these, and were given three examples of novel names these kinds of shapes might have. The object shapes, colors, and names in this introduction phase were randomly chosen and different from the ones used in the task.

In this one-trial task, participants saw a stimulus display with six identical block shapes (distractors) surrounding one unique block shape (the target) (see label in Figure ??). They were asked to “Find the [blue] toma,” with an utterance that either included a color adjective or did not. We expected

that participants would click on the unique, central object, and they only moved forward in the task once they selected that object.

Two conditions of interest, utterance type (with or without an adjective) and display type (whether the distractors’ color was the same as or different from the target) were varied between subjects. Additionally, several factors were randomly assigned between subjects: the name of the target (among *dax*, *blicket*, *wug*, *toma*, *gade*, or *sprock*), the target color (among red, blue, green, orange, purple, yellow, pink, tan, teal, or grey), the distractor color if different from the target (among the same set of colors), and the target and distractor shapes (among ten possible block shapes with five square components).

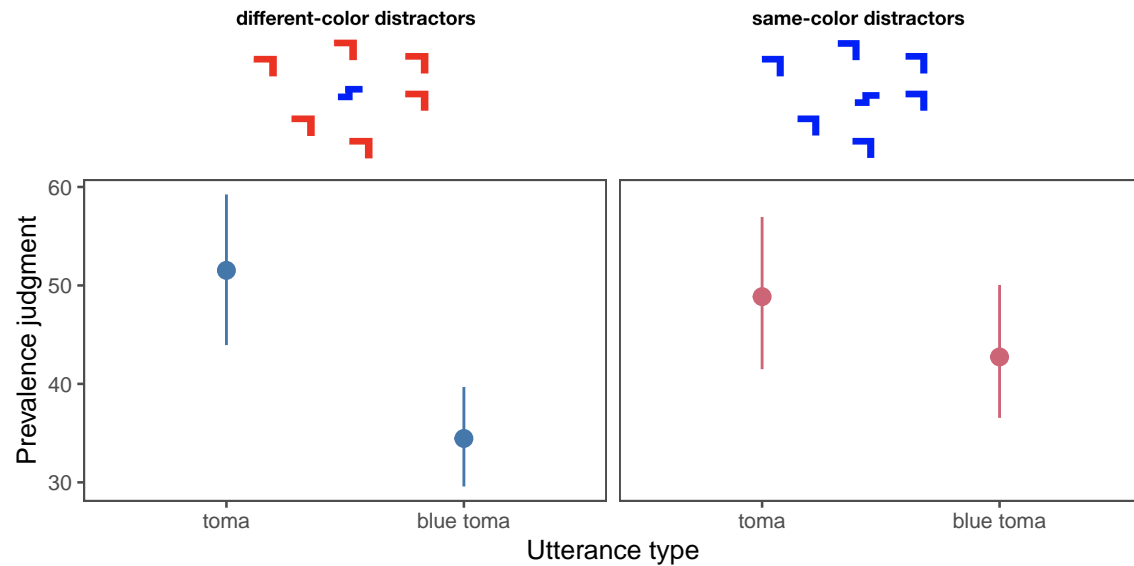
After selecting the target object, participants made a judgment about the prevalence of the target’s color among the target’s category. They were asked, e.g., “What percentage of tomas do you think are blue?” and responded on a slider scale between 0 and 100%.

S6.0.3 Results

The main question of interest is whether people infer that a color is rarer when it is mentioned than when it is not; when a participant hears “blue toma,” do they infer that tomas are less likely to be blue? We fit a linear model predicting participants’ prevalence judgments from the utterance type (noun or adjective noun) and context type (different-color or same-color distractors) and their interaction. There was a significant effect of utterance type, such that people’s prevalence judgments were lower when there was an adjective in the utterance than when there was not ($\beta_{adjective} = -17.07$, $t = -3.65$, $p = < .001$). There was not a significant effect of context type ($\beta_{same-color-context} = -2.67$, $t = -0.54$, $p = .587$) or an utterance by context interaction ($\beta_{adjective-context} = 10.95$, $t = 1.61$, $p = .108$).

S6.1 Discussion

We found that when a block object is referred to as “the blue toma” rather than “the toma,” people think tomas are less likely to be blue in general. People’s judgments also did not significantly differ depending on whether surrounding objects were the same color as the target. This stripped-down demonstration of the effect provides additional evidence that people infer described objects are atypical even when they likely do not have strong expectations that the object categories have a stereotypical color (as they may with fruit). It is an open and interesting question whether people’s prior expectations about a category’s feature distribution would modulate or extinguish this effect. This demonstration is evidence that the effect generalizes beyond fruit; we leave a systematic investigation of that question to future work.



References

- Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to “overinformative” referring expressions. *Psychological Review*, 127, 591–621.
- Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, 27(1), 89–110.
- Rubio-Fernández, P. (2016). How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology*, 7.