Stimuli, model code, model predictions, data, and analyses scripts are available at: https://osf.io/bezua/

Model validation studies

Regression tables for Model 1: Percentage of dwell time on the Target

lmer(PERCENTAGE_LOOKS ~ Condition + (1 + Condition | Subject) + (1 + Condition | Item))

Table 1: Fixed effects

	Estimate	**Standard Error**	***t* value**
Intercept	37.619	2.015	18.673
Color condition	3.911	1.142	3.423

Table 2: Random effects

Groups	**Name**	**Variance**	**St Deviation**	**Correlation**
Subject	Intercept	69.704	8.349	
	Color condition	2.848	1.688	1.00
Item	Intercept	43.500	6.595	
	Color condition	12.652	3.557	0.08
Residual		165.681	12.872	

Regression tables for Model 2: Percentage of dwell time on the Contrast object

lmer(PERCENTAGE_LOOKS ~ Condition + (1 + Condition | Subject) + (1 + Condition | Item))

Table 1: Fixed effects

	Estimate	**Standard Error**	***t* value**

Intercept	24.392	1.625	15.013
Color condition	-7.039	1.213	-5.801

Table 2: Random effects

Groups	Name	Variance	St Deviation	Correlation
Subject	Intercept	24.357	4.935	
	Color condition	2.667	1.633	-1.00
Item	Intercept	47.230	6.872	
	Color condition	24.901	4.990	-0.03
Residual		116.074	10.774	

Regression tables for Model 3: Response times

lmer(RT ~ Condition + (1 + Condition | Subject) + (1 + Condition | Item))

Table 1: Fixed effects

	Estimate	Standard Error	*t* value
Intercept	2187.22	44.74	48.892
Color condition	-402.37	41.61	-9.669

Table 2: Random effects

Groups	Name	Variance	St Deviation	Correlation
Subject	Intercept	39641	199.1	

	Color condition	21569	146.9	-0.24
Item	Intercept	18113	134.6	
	Color condition	23006	151.7	-0.81
Residual		59848	244.6	

Model implementation details

Alignment of model parameters

Analysis code is available on our OSF site. As reported in the main text, we began by setting our ICE model parameters prior to data collection and we then found the Brevity model parameters that made its range of predictions closest to the ICE model. That is, we sought to find a Brevity model whose distribution of confidence matched that of the ICE model (although the confidence would occur in different trials).

To achieve this, we performed the following operation: for each set of parameters that we considered (word cost values from 0.01 to 1 in jumps 1, and tau values from 0.01 to 3 in jumps of 0.01, for a total of 30,000 parameter combinations), we generated the Brevity model predictions for our Experiment and sorted the predictions in ascending order (i.e., combining all predictions from all trials). We next also sorted the predictions from the ICE model (with pre-set parameters), and we computed the Euclidean distance between each pair of points. Thus, the lower the Euclidean distance, the more the distribution of judgments was aligned.

Model fitting

As reported in the main text, our main evaluations used pre-set parameters. For completeness we also fit each model's parameters to maximize model fit. Doing so further revealed that the alternative Brevity model was significantly worse than the ICE model, both when using the pre-set parameters and the best fit.

Note that for both models, the reward associated with communicating successfully was set as a constant value (set to $R(t)=30$ for the ICE model and $R(t)=1$ for the Brevity model). The purpose of this reward was only to ensure that the models were always sufficiently motivated to communicate unambiguously. We thus used a higher value for the ICE model because the range of costs was higher relative to the range of costs from the Brevity model. In principle, we could also fit these values during the parameter search. Note, however, that the models make the final decision based on the utility—the difference between the costs and the rewards. Thus,

the critical dimension that affects model fit is the difference between these two variables. Consequently, leaving the reward fixed and varying the costs produces the same effect as varying both variables and we therefore left the rewards fixed during the parameter search process.

For our ICE model, we searched the combinatorial space for FPW values between 1 and 9, and tau values between 0.01 and 3. We initially began searching for the best tau parameter using 0.01 steps (i.e., 0.01, 0.02, 0.03, …). However, due to computational feasibility, we stopped this search at 0.12, and restarted it with jumps 0.1 (beginning a 0.2, for a total of 40 tau values). For each tau proposed tau parameter, we tested every integer FPW between 1 and 9 (9 parameters total). Combined, this led to 360 (9*40) parameter combinations tested, with the best fit at parameters FPW=1 and tau=2.9.

For the alternative Brevity model, we searched the combinatorial space for word costs between 0.01 to 1 in jumps of 0.01 (100 parameters total), and tau values between 0.01 and 3 in jumps of 0.01 (300 parameters). Combined, this led to 30,000 (100*300) parameter combinations tested, with the best fit at word cost = 0.01 and tau = 0.31.

Note that while the parameter search for the Brevity model was more exhaustive relative to the ICE model, this asymmetry can only bias the results *against* our account, as we tested over 80 times more parameter combinations for the Brevity model, giving it a higher opportunity to find a combination that could outperform our ICE model.

Experiment Supplemental Results

Scatterplot using best model fits

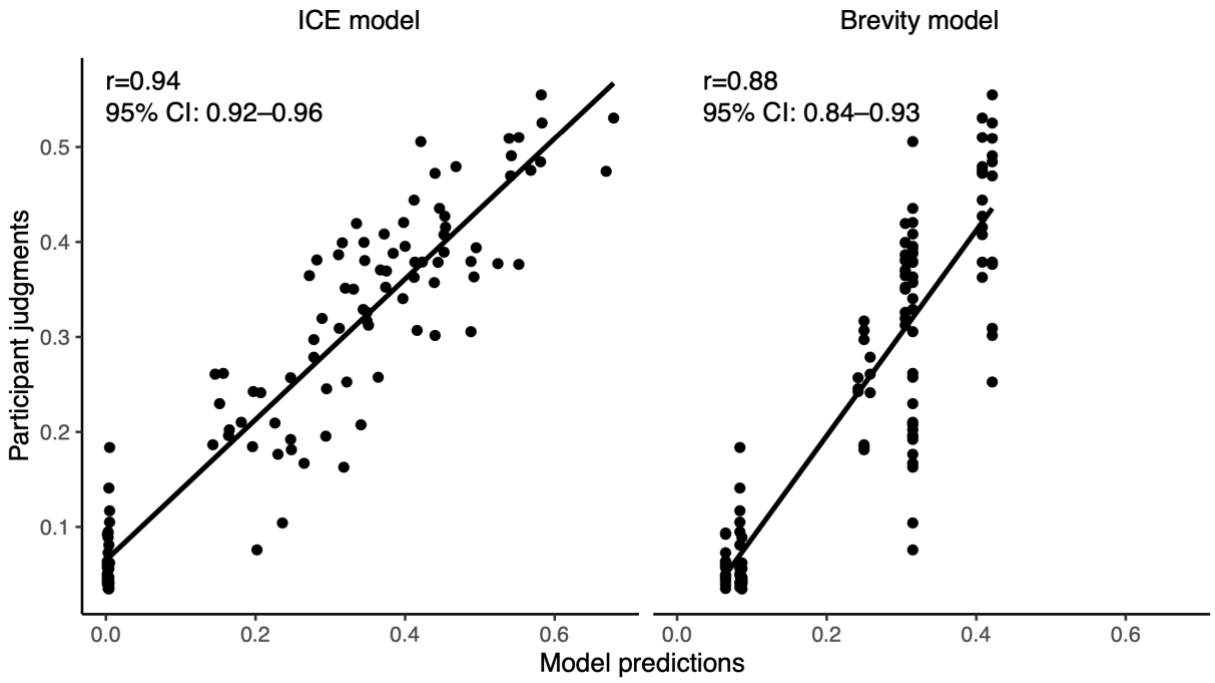

Fig. S1. Model predictions (x axis) plotted against participant judgments (y axis) using the parameters that maximized each model's correlation.

For trial-by-trial predictions please see our OSF repository.

Speed of reference production

Because our model uses Monte Carlo estimates to calculate the listener's visual search, it is possible to quantify how many samples are needed before our model is converges to its final answer. We can therefore use the number of samples as a proxy for the time it takes for our model to decide what to say in different trials.

To test if our model is slower as a function of objects in a scene, we generated artificial visual displays with 2, 4, 8, and 16 objects. In each of these displays the object had a unique color, material, and category, such that any expression would resolve the referent. We next calculated the utility of each possible utterance using 1000 Monte Carlo simulations and treated the final utility estimates as ground truth.

Using these final utility estimates, it is possible to set an error threshold, and compute at which point in the sampling process all of the utility estimates fall below the threshold:

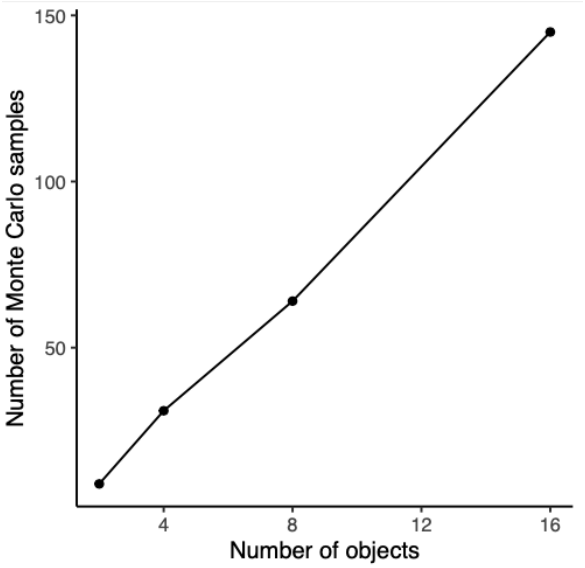

Fig. S2. Our model shows it is more difficult to decide which utterance to use, as a function of the number of objects in a scene. The x axis shows number of objects in the scene and the y axis shows how many Monte Carlo samples are needed for the utility estimates of each utterance to be at most 0.5 off from the final utilities (the pattern of results is qualitatively identical under different error thresholds).

This pattern is qualitatively similar to the one found by Gatt et al. (2017), where speakers are slower to generate a referential expression as a function of scene complexity.