Supplementary information:Pragmatic cue integration in adults' and children's inferences about novel word meanings

Manuel Bohn[1,2], Michael Henry Tessler[3], Megan Merrick[1], & Michael C. Frank[1]

[1] Department of Psychology, Stanford University
[2] Leipzig Research Center for Early Child Development, Leipzig University
[3] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Overview

Here we present details for the cognitive models as well as supplementary analysis and results. Readers who are interested in the model code and analysis code itself are encouraged to consult the associated online repository: https://github.com/manuelbohn/mcc.

Cognitive Models

Cognitive models were implemented in WebPPL (Goodman & Stuhlmüller, 2014) using the r package `rwebppl` (Braginsky, Tessler, & Hawkins, 2019).

Ontology

The situation we model is defined by three sets: referents r, utterances u, and lexica $\mathcal{L}$. Referents are defined by two features $r_{t,l}$: the type of object they are t (visually discernible given their shape and color) and their location l (one of two tables). There are two types of objects t_1, t_2 and two locations l_1 and l_2. The utterances u available to a speaker are (action, label) pairs (u_a, u_w), where the action involves turning (pointing) to a particular location (a table) and labels are novel words (w_1, w_2), which are assumed to refer to the *type* of the object. There are two kinds of lexica, corresponding to the two utterance components: The lexicon of pointing $\mathcal{L}_{point}$ and the lexica of labels $\mathcal{L}_1, \mathcal{L}_2$. The lexicon of pointing $\mathcal{L}_{point}$ encodes the meaning of a point: The action of pointing reduces the set of referents to those that are on the table targeted by the point. The lexica of labels map labels u_w onto referents: $\mathcal{L}_1$ which maps label 1 w_1 to type 1 t_1 and label 2 w_2 to type 2 t_2, and $\mathcal{L}_2$ which has the inverse mapping.

Pragmatics model

Our word-learning model is a model of pragmatic reasoning couched in the Rational Speech Act modeling framework (Frank & Goodman, 2012; Goodman & Frank, 2016). The model describes the following process: A pragmatic listener (L_1) jointly infers a referent

(what object is being picked out by the utterance) and a lexicon (label–type mappings) by reasoning about a pragmatic speaker (S_1) who produces utterances to convey information to a literal listener (L_0), who in turn interprets utterances according to their literal meaning. According to Bayes rule, the pragmatic listener's inference is given by:

$$P_{L_1}(r, \mathcal{L}|u) \propto P_{S_1}(u|r_t, \mathcal{L})P(\mathcal{L})P(r) \tag{1}$$

The right-hand side of this equation has three terms: the prior distribution over referents $P(r)$, the prior distribution over lexica $P(\mathcal{L})$, and the likelihood $P_{S1}(u|r_t, \mathcal{L})$ that a speaker would produce an utterance u given a referent type r_t and their lexicon $\mathcal{L}$.

In the situation we model, the prior on referents $P(r)$ is a categorical distribution over three objects in a scene, which we posit could be non-uniform due to what is in common ground (see main text and below for information on common ground manipulations). Because the labels produced by the speaker are all novel words, the listener has no substantive knowledge about the lexica (label–type mappings) and thus the prior over the two lexica (described above) is uniform.

The pragmatic listener updates their beliefs about both the referent and the lexicon by reasoning about the speaker, assumed to produce utterances to convey the referent type (r_t) to the listener by being a soft-max rational agent (with degree of rationality α) with a utility function defined in terms of the informativity of an utterance for a referent type (r_t):

$$P_{S_1}(u_{a,l}|r_t, \mathcal{L}) \propto Informativity(u_a; r_t)^{\alpha} \cdot P(u_l \mid \mathcal{L}) \tag{2}$$

The speaker's utterances are (action, label) pairs (u_a, u_l) (i.e., we assume the speaker must point to one of the locations and must produce a label). The label the speaker produces u_l depends upon their lexicon $\mathcal{L}$ (i.e., they choose the label that is consistent with the object under their lexicon).The action (point) the speaker produces u_a is a function of that signal's informativity. That is, because the listener does not know the meaning of the labels, the labels carry no information to the listener.

The informativity of the utterance (action) for a referent is the probability that a naive listener L_0 would select a the type of referent (r_t) given that utterance (action)(Goodman & Stuhlmüller, 2013).

$$Informativity(u_a; r_t) = P_{L_0}(r_t|u_a) \tag{3}$$

where the probability that the naive listener L_0 would select a referent of a given type r_t given the utterance u_a is given by Bayes' Rule:

$$P_{L_0}(r_t|u) = \sum_{r_i \in t} P_{L_0}(r_i|u) \propto \mathcal{L}_{point} P(r_i) \tag{4}$$

where $\mathcal{L}_{point}$ encodes the meaning of the actions (points), and is simply a truth-function stipulating that the location of the referent must be at the location of the point.

We assume in this model that the speaker is trying to convey the type of the referent r_t rather than the individual token referents r_i; thus, the relevant probability for purposes of informativity is the type probability $P_{L_0}(r_t|u)$, which is simply a sum of the token probabilities for tokens that are of the same type $\sum_{r_i \in t} P_{L_0}(r_i|u)$.

Figure S1. Screenshot showing test situation Experiment 1, 3 and 4.

$P(r)$ again denotes the prior probability of a referent, $\mathcal{L}_{point}$ encodes the literal meaning of a point, returning 1 if the object is at the location of the point and 0 if the utterance if not. As mentioned above, because L_1 does not know the lexicon, the semantics of the words contained in the utterance (i.e., the labels) offer no information about the referent. The non-linguistic aspect of the utterance (pointing), however, do. As described above, the semantics of turning to one of the tables is roughly equivalent to saying "It's an object on that table".

The above formulation of a literal interpretation model based only on the semantics of the non-linguistic signals (points) can also be derived by positing a literal interpretation model that updates their beliefs according to the literal meanings of both labels and points, but which is uncertain about the meaning of the labels: $P_{L_0}(r, \mathcal{L}|u) \propto \mathcal{L}_{lit}P(r)P(\mathcal{L})$.

Worked Example

In this section, we work through a toy numerical example for how model predictions were generated for the pragmatics model. The prediction will correspond to the parameter free model described above (excluding the noise parameter). The values of the parameters are taken from a preference condition (see below). Fig. S1 shows a screenshot from the adult experiment, which the model was designed to capture. In this context, there are three potential referents of two types (pink-ish and yellow-ish; for simplicity, we refer refer the object's type by its color) on two tables: $r_{type:pink,table:1}$, $r_{type:pink,table:2}$; one yellow-ish, $r_{type:yellow,table:2}$). For simplicity, we refer to the referents by the shorthand: r_{p1}, r_{p2}, r_{y2}. In principle, the speaker, the frog in this case, can produce one of four utterances, pointing either to the left or right table and saying either label ("dax" or "wug"): $u_1 = (\text{left}, dax)$, $u_2 = (\text{left}, wug)$, $u_3 = (\text{right}, dax)$, $u_4 = (\text{right}, wug)$. We work out the example where the speaker produces utterance u_3, pointing to the right table (two objects) and saying the label "dax" (though since the labels have no *a priori* meanings, the computation would be the same for utterance u_4).

As mentioned above, the listener is learning the mappings between labels and object types (rather than object tokens). That is, the listener either believes the novel word "dax" refers to "pink-ish objects" or "yellow-ish objects". Pointing to a table always has the same meaning. These semantics can be described using two lexica: $\mathcal{L}_1 = \{dax : \text{pink-ish thing}, wug : \text{yellow-ish thing}, point : \text{location of point}\}$, $\mathcal{L}_2 = \{dax : \text{yellow-ish thing}, wug : \text{pink-ish thing}, point : \text{location of point}\}$.

We construct the prior distribution over referents $P(r)$ based on the results of Experiment 2A, in which a different speaker displayed a preference for a yellow-ish object. The prior distribution over referents $P(r)$ (left to right in Fig. S1) in the example that follows was [0.26, 0.26, 0.48] (see Model Parameters section below for a detailed description of how this distribution was constructed).

We assume that the intentional goal of the speaker is to get the listener to select an object of the correct type. That is, the informativity of an utterance is calculated with respect to conveying the correct object type as opposed to a particular referent (token).

First, we calculate the literal listener's posterior distribution over referents, and marginalize (average) over objects of the same type to compute the informativity of an utterance for a type.

$$
\begin{aligned}
P_{L_0}(r_{p1}|u_3) &\propto \mathcal{L}_{point}(r_{p1}, u_3)P(r_{p1}) = 0 \times 0.26 = 0 \\
P_{L_0}(r_{p2}|u_3) &\propto \mathcal{L}_{point}(r_{p2}, u_3)P(r_{p2}) = 1 \times 0.26 = 0.26 \\
P_{L_0}(r_{y2}|u_3) &\propto \mathcal{L}_{point}(r_{y2}, u_3)P(r_{y2}) = 1 \times 0.48 = 0.48
\end{aligned}
$$

because the speaker is pointing to the right table with r_{p2} and r_{y2}. After normalization we have:

$$
\begin{aligned}
P_{L_0}(r_{p1}|u_3) &= 0 \\
P_{L_0}(r_{p2}|u_3) &= 0.35 \\
P_{L_0}(r_{y2}|u_3) &= 0.65
\end{aligned}
$$

In order to get the distribution over types (instead of tokens), we simply add up the probabilities for each token from the same type.

$$
\begin{aligned}
Informativity(u_3; r_p) &= P_{L_0}(r_p|u_3) = P_{L_0}(r_{p1}|u_3) + P_{L_0}(r_{p2}|u_3) = 0 + 0.35 = 0.35 \\
Informativity(u_3; r_p) &= P_{L_0}(r_y|u_3) = P_{L_0}(r_{y2}|u_3) = 0.65
\end{aligned}
$$

In a similar way, we can compute the informativity of the other utterances. Since the words have no *a priori* meanings, u_4 (pointing to the right and saying "wug") will have the same informativity values as u_3 (pointing the right and saying "dax"). The informativity vectors for u_1 and u_2 (pointing to the left and saying either "dax" or "wug") are also identical and given by:

$$
\begin{aligned}
Informativity(u_1; r_p) &= P_{L_0}(r_p|u_1) = 1 \\
Informativity(u_1; r_y) &= P_{L_0}(r_y|u_1) = 0
\end{aligned}
$$

We assume that the speaker knows the lexicon, but doesn't believe the listener to know the lexicon. That is, the generative process of the utterance is tantamount to pointing to the

table with the referent of the type the speaker wants and incidentally labeling it. The label itself carries no information.

Based on equation 2 we can now compute the likelihood of each utterance given an object type. Recall the two lexica: $\mathcal{L}_1 = \{dax : \text{pink-ish thing}, wug : \text{yellow-ish thing}\}$, $\mathcal{L}_2 = \{dax : \text{yellow-ish thing}, wug : \text{pink-ish thing}\}$, and the four utterances: $u_1 = (\text{left}, dax)$, $u_2 = (\text{left}, wug)$, $u_3 = (\text{right}, dax)$, $u_4 = (\text{right}, wug)$. We begin with the pink-ish type (r_p). If the speaker's lexicon were $\mathcal{L}_1$:

$$P_{S_1}(u|r_p, \mathcal{L}_1) \propto Informativity(u_a; r_p)^\alpha P(u \mid \mathcal{L}_1) = \begin{cases} 1^{2.24} \times 1 = 1 & \text{for } u_1 \\ 0^{2.24} \times 0 = 0 & \text{for } u_2 \\ 0.35^{2.24} \times 1 = 0.095 & \text{for } u_3 \\ 1^{2.24} \times 0 = 0 & \text{for } u_4 \end{cases}$$

To arrive at the production probabilities, we normalize by the values from the previous calculation so that they add to 1:

$$P_{S_1}(u|r_p, \mathcal{L}_1) = \begin{cases} 0.91 & \text{for } u_1 \\ 0 & \text{for } u_2 \\ 0.09 & \text{for } u_3 \\ 0 & \text{for } u_4 \end{cases}$$

The speaker's production probabilities are the same if the speaker's lexicon is $\mathcal{L}_2$, only the utterances use the other label:

$$P_{S_1}(u|r_p, \mathcal{L}_2) = \begin{cases} 0 & \text{for } u_1 \\ 0.91 & \text{for } u_2 \\ 0 & \text{for } u_3 \\ 0.09 & \text{for } u_4 \end{cases}$$

That is, if the speaker is trying to convey the pink-ish object type, the speaker would be 10 times more likely (under a speaker rationality parameter of 2.24) to point to the left table than to the right table. This corresponds with the intuition that pointing to the left table would be a much better way to refer to a pink-ish object. If instead the speaker wanted to convey the yellow-ish object type (r_y):

$$P_{S_1}(u|r_y, \mathcal{L}_1) = Informativity(u; r_y)^\alpha \cdot P(u \mid \mathcal{L}_1) = \begin{cases} 0^{2.24} \times 0 = 0 & \text{for } u_1 \\ 0^{2.24} \times 0 = 0 & \text{for } u_2 \\ 0^{2.24} \times 0 = 0 & \text{for } u_3 \\ 0.65^{2.24} \times 1 = 0.38 & \text{for } u_4 \end{cases}$$

To arrive at the production probabilities, we normalize:

$$P_{S_1}(u|r_y, \mathcal{L}_1) = \begin{cases} 0 & \text{for } u_1 \\ 0 & \text{for } u_2 \\ 0 & \text{for } u_3 \\ 1 & \text{for } u_4 \end{cases}$$

and

$$P_{S_1}(u|r_y, \mathcal{L}_2) = \begin{cases} 0 & \text{for } u_1 \\ 0 & \text{for } u_2 \\ 1 & \text{for } u_3 \\ 0 & \text{for } u_4 \end{cases}$$

That is, the speaker would point to the right table and say the utterance consistent with their lexicon. Again, intuitively this makes sense because the yellow-ish object is located only on the right table and pointing to that table presents the only way one could refer to that object type.

Finally, based on equation 1 we can use these values to compute the probability that the listener thinks that the speaker is referring to the yellow-ish type (r_y) when they produce u_3 (pointing to the right table and saying "dax"). In this case, it is the same as the probability of the yellow-ish referent (r_{y2}) because there is only one yellow object.

$$P_{L_1}(r_{y2}, \mathcal{L}|u_3) = \frac{P_{S_1}(u_3 \mid r_y, \mathcal{L}) \cdot P(\mathcal{L}) \cdot P(r_{y2})}{\sum_{r'} P_{S_1}(u_3 \mid r', \mathcal{L}) \cdot P(\mathcal{L}) \cdot P(r')}$$

$$\begin{aligned}
P_{L_1}(r_{p1}, \mathcal{L}_1|u_3) &\propto P_{S_1}(u_3 \mid r_{p1}, \mathcal{L}_1)P(r_{p1})P(\mathcal{L}_1) = 0.09 \times 0.26 \times 0.5 = 0.012 \\
P_{L_1}(r_{p2}, \mathcal{L}_1|u_3) &\propto P_{S_1}(u_3 \mid r_{p2}, \mathcal{L}_1)P(r_{p2})P(\mathcal{L}_1) = 0.09 \times 0.26 \times 0.5 = 0.012 \\
P_{L_1}(r_{y2}, \mathcal{L}_1|u_3) &\propto P_{S_1}(u_3 \mid r_{y2}, \mathcal{L}_1)P(r_{y2})P(\mathcal{L}_1) = 0 \times 0.48 \times 0.5 = 0 \\
P_{L_1}(r_{p1}, \mathcal{L}_2|u_3) &\propto P_{S_1}(u_3 \mid r_{p1}, \mathcal{L}_2)P(r_{p1})P(\mathcal{L}_2) = 0 \times 0.26 \times 0.5 = 0 \\
P_{L_1}(r_{p2}, \mathcal{L}_2|u_3) &\propto P_{S_1}(u_3 \mid r_{p2}, \mathcal{L}_2)P(r_{p2})P(\mathcal{L}_2) = 0 \times 0.26 \times 0.5 = 0 \\
P_{L_1}(r_{y2}, \mathcal{L}_2|u_3) &\propto P_{S_1}(u_3 \mid r_{y2}, \mathcal{L}_2)P(r_{y2})P(\mathcal{L}_2) = 1 \times 0.48 \times 0.5 = 0.24
\end{aligned}$$

After normalizing, the full joint-posterior distribution over referents and lexica is:

$$\begin{aligned}
P_{L_1}(r_{p1}, \mathcal{L}_1|u_3) &= 0.045 \\
P_{L_1}(r_{p2}, \mathcal{L}_1|u_3) &= 0.045 \\
P_{L_1}(r_{y2}, \mathcal{L}_1|u_3) &= 0 \\
P_{L_1}(r_{p1}, \mathcal{L}_2|u_3) &= 0 \\
P_{L_1}(r_{p2}, \mathcal{L}_2|u_3) &= 0 \\
P_{L_1}(r_{y2}, \mathcal{L}_2|u_3) &= 0.91
\end{aligned}$$

To arrive at the distribution over types, you simply add all of the referents of the same type and marginalize out the lexicon:

$$P_{L_1}(r_p|u_3) = \sum_{i \in 1,2} P_{L_1}(r_{p1}, \mathcal{L}_i|u_3) + P_{L_0}(r_{p2}, \mathcal{L}_i|u_3) = 0.09$$
$$P_{L_1}(r_y|u_3) = \sum_{i \in 1,2} P_{L_1}(r_{y2}, \mathcal{L}_i|u_3) = 0.91$$

To conclude the example, according to the parameter free pragmatics model, the probability that L_1 thinks that S_1 is referring to a yellow-ish object when producing u_3 is 0.91. Based on this model we thus expected that, on average, participants would select the yellow-ish object in 91% of cases in a condition with the prior distribution specified above.

Prior only model

The prior only model ignored the information about the intended referent that was expressed by the utterance and instead only focused on common ground manipulation. The only information available to L_1 is the prior distribution over referents. It is therefore defined as:

$$P_{L_1}(r|u) \propto P(r) \tag{5}$$

That is, the probability of the referent given the utterance is determined by the prior probability of the referent for a particular speaker. The prior distributions were set in the same way as for the pragmatics model.

Flat prior model

This model was identical in structure to the pragmatics model with the exception that the prior distribution did not correspond to the measurements from Experiment 2 and did not vary with speaker identity. That is, regardless of common ground manipulation and speaker identity the prior distribution was always uniform (i.e. [0.33,0.33,0.33]). The speaker optimality parameter α was set in the same way as in the pragmatics model.

Model parameters

As noted in the main text, the parameter α (speaker optimality parameter) in equation 2 determines the absolute strength of the likelihood term. It's interpretation is *how* rational L_1 thinks S_1 is in this particular context. For adults, we used the data from Experiment 1 to infer the value of α. That is, we inferred which value of α would generate model predictions for the pragmatics model (assuming equal prior probability over referents) that corresponded to the average proportion of correct responses measured in Experiment 1. This value for α was then used in Experiment 3 and 4.

For children, the speaker optimality parameter changed with age. Instead of inferring a single value across age, we used the data from Experiment 5 to find the slope and intercept for α that best described the developmental trajectory in the data. As for adults, this was done via the pragmatics model with equal prior probability for each object. In Experiment 7, the speaker optimality parameter for a given child of a given age was computed by taking the overall intercept and adding the slope times the child's age (with age anchored at 0).

Figure S2. Screenshot showing test situation in Experiment 2 for adults.

The analysis code corresponding to these calculations can be found in the associated online repository.

The prior distribution over objects, $P(r)$, varied with the common ground manipulation, the identity of the speaker and the alignment of utterance and common ground information. Numerically, it depended on the measurement obtained in Experiment 2A and B for adults and Experiment 6 for children. Fig. S2 shows the test situation on Experiment 2.

For adults, this worked in the following way: For example, in Experiment 2, for the preference/same speaker condition, when the speaker previously indicated that they liked the yellow-ish object (right table in Fig. S2) and disliked the pink-ish object (left table in Fig.S2), the average proportion with which participants chose the yellow-ish object was 0.97 and for the pink-ish object it was 0.03 respectively. In Experiment 3 and 4, this measurement determined the prior distribution over objects in cases whenever the the same manipulation was used (preference/same speaker). Note that Experiment 3 involved three objects while Experiment 2 only involved two. We nevertheless used the exact proportions measured in Experiment 2 for each object to inform the prior. This approach spread out the absolute probability mass but conserved the relative relation between objects. Thus, when utterance and common ground information were aligned (i.e. the yellow-ish object was the more informative object as in Fig. S1), the distribution of objects was $[r_{p1}, r_{p2}, r_{y2}]$. Using the raw proportions, the corresponding prior distribution was $[P(r_p^1) = 0.03, P(r_p^2) = 0.03, P(r_y^2) = 0.97]$ and after normalizing (so that they add up to 1) it was [0.03, 0.03, 0.94]. When information sources were dis-aligned (i.e. the pink-ish object was the more informative one), the object distribution was $[r_{y1}, r_{p2}, r_{y2}]$ and the prior distribution was thus [0.97, 0.03, 0.97] or [0.49, 0.02, 0.49] after normalizing.

For children, we used the data from Experiment 6 to model the slope and intercept that best described the developmental trajectory in the data for each of the two conditions. As for the speaker optimality parameter, this allowed us to generate prior distributions that were sensitive to the child's age. In Experiment 7, the prior probability for an object was

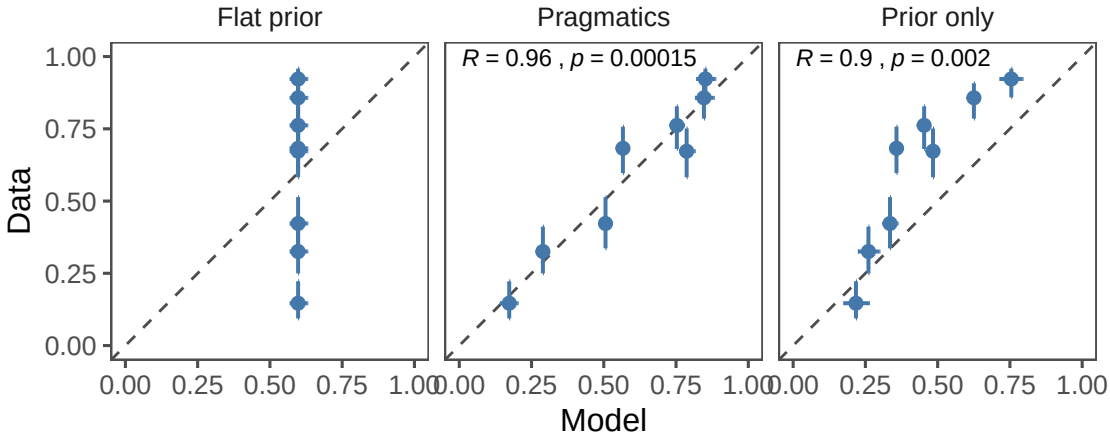

Figure S3. Correlation plot for model predictions and data from Experiment 3. All models depicted here included a noise parameter. Coefficients and p-values are based on Pearson correlation statistics. Dots represent condition modes. Error bars represent 95% HDIs.

computed by taking the intercept for the respective condition (same or different speaker), adding the slope times the child's age and then using a logistic transformation to convert the outcome into proportions. The overall distribution then depended on the alignment of information sources in the same way as it did for adults. Model code for inferring the intercept and the slope for the child study can be found in the associated online repository.

Model comparison

Analysis code for model comparison can be found in the online repository.

Experiment 3

Here we report details on the model comparisons. Model fit was assessed based on marginal log-likelihoods of the data under each model. Bayes Factors were computed by first subtracting log-likelihoods for two models and then exponentiating the result. Table S1 shows Bayes Factors for model comparisons in Experiment 3. We did not pre-register the inclusion of the noise parameter for Experiment 3, but did so for all subsequent experiments for which we did model comparisons (4 and 7). The first row in Table S1 compares the pragmatics model with noise parameter to the model without the noise parameter. This comparison shows that including the noise parameter greatly improves model fit. Figure S3 correlates model predictions from the models including noise parameters to the data from Experiment 3.

Figure S4 shows the posterior distribution of the noise parameter under each model. The noise parameter was fit to the data and indicates the proportion of responses that are estimated to be due to random guessing rather than in line with model predictions. Consequently, a model that makes predictions that are closer to the data is likely to have a lower noise parameter. The results corroborate the model comparison by showing that the pragmatics model has the lowest noise parameter.

Table S1
Bayes Factors for model comparisons in Experiment 3

Comparison	BF
pragmatic_noise > pragmatic	2.2e+295
pragmatic_noise > prior_only_noise	2.5e+34
pragmatic_noise > flat_prior_noise	4.2e+53
prior_only_noise > flat_prior_noise	1.7e+19

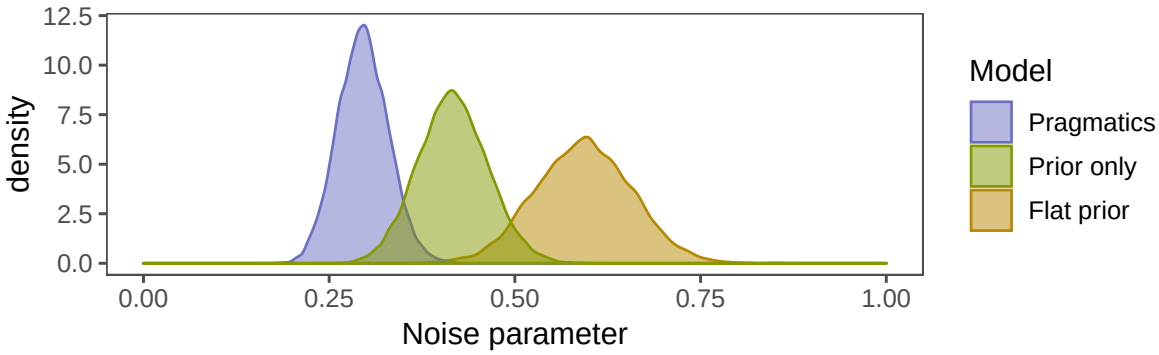

Figure S4. Posterior distribution of noise parameter for each model in Experiment 4

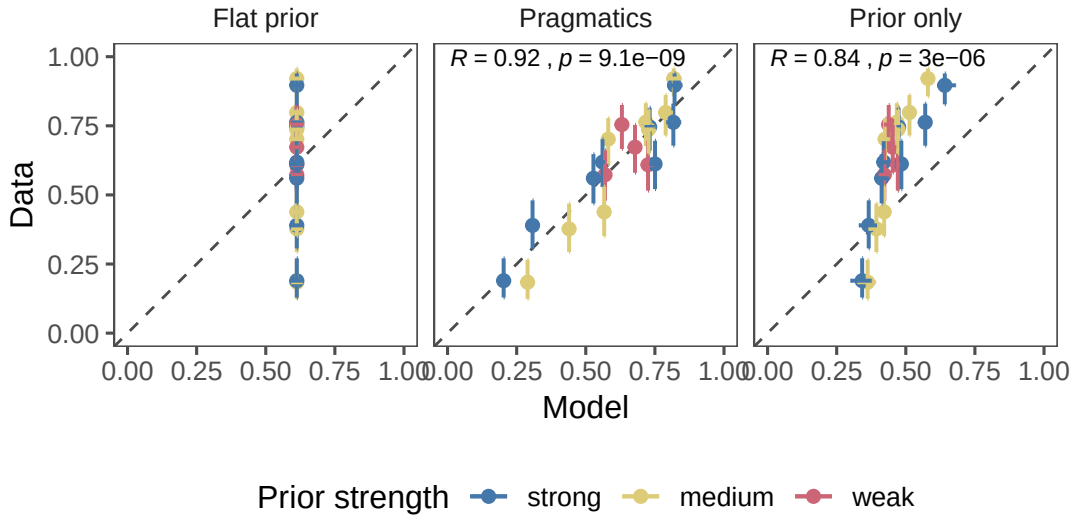

Figure S5. Correlation plot for model predictions and data from Experiment 4. All models included a noise parameter. Coefficients and p-values are based on Pearson correlation statistics. Dots represent condition modes. Error bars represent 95% HDIs.

Table S2
Model comparisons in Experiment 4

Comparison	BF
Pragmatics > Prior only	8.9e+82
Pragmatics > Flat prior	4.7e+71
Flat prior > Prior only	1.9e+11

Note. BF = Bayes Factor; All models include a noise parameter.

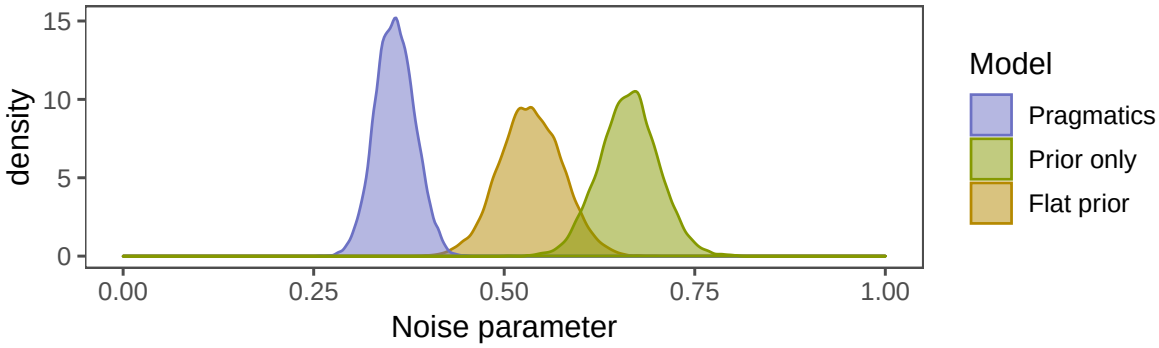

Figure S6. Posterior distribution of noise parameter for each model in Experiment 4

Experiment 4

Figure S5 correlates model predictions with the data from Experiment 4. Table S2 shows Bayes Factors for model comparisons in Experiment 4. As pre-registered, all models included a noise parameter. Figure S6 shows the posterior distribution of the noise parameter for each model in Experiment 4. All results suggest that the pragmatics model captures the structure in the data better compared to the alternative models considered.

Experiment 7

For children, we compared models using different types of noise parameters. We preregistered the model comparison for models including a single noise parameter. We added the additional model comparisons because the noise parameter was relatively high. The additional model comparisons allow us to see if the pragmatics model provides a better fit when more emphasis is put on the model structure itself. The results show that this was the case.

Parameter free models did not include a noise parameter. Noise models included a single noise parameter for all ages. Developmental noise models included a noise parameter that changed with age. That is, instead of a single value, we inferred an intercept and a slope for the noise parameter. Noise was therefore a function of the child's age. Table S3 shows model comparisons for the pragmatics models using different noise parameters. This shows that including a noise parameter improves model fit but that the type of noise parameter does not make much of a difference.

Table S3
Model comparisons for pragmatics models in Experiment 7

Comparison	BF
dev. noise > noise	1.5
noise > parameter free	1.1e+03
dev. noise > parameter free	1.6e+03

Note. BF = Bayes Factor

Table S4
Model comparisons in Experiment 7

Parameter	Pragmatics > Flat P.	Pragmatics > P. only	Flat P. > P. only
developmental noise	1.6e+04	1e+06	63
noise	5.8e+02	1.1e+04	18
parameter free	3.3e+02	20	0.06

Note. BF = Bayes Factor

Table S4 shows results for model comparison for the different types of noise parameters. In all cases, the pragmatics model provides a substantially better fit to the data compared to the alternative models.

Figure S7 shows the different types of noise parameters for the each model. Figure S7A shows that the pragmatics model has the lowest estimated level of noise of all the models considered. Figure S7B shows that the the pragmatics model has the lowest level of estimated noise across the entire age range. It also shows that noise decreases with age for the pragmatics model, suggesting that older children behaved more in line with model predictions compared to younger children.

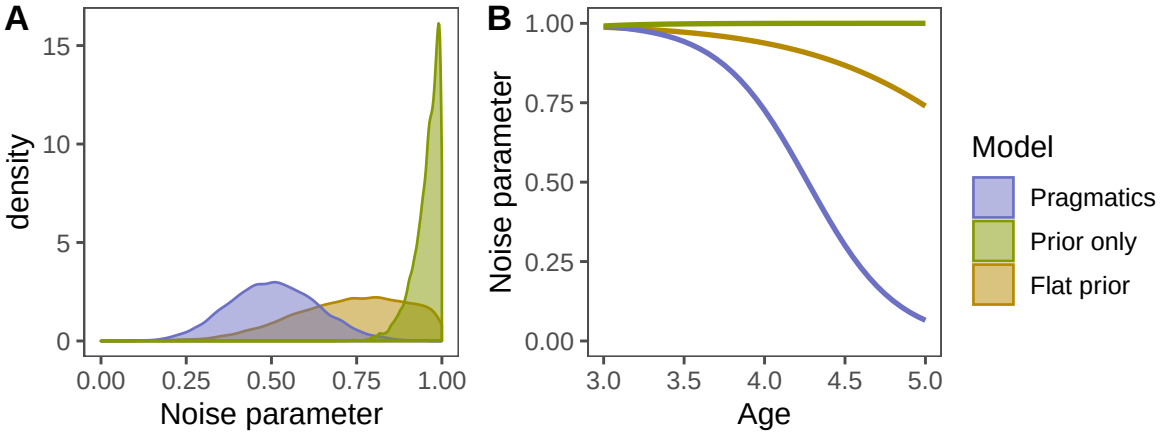

Figure S7. Posterior distribution of noise parameter for each model in Experiment 7. A: single noise parameter across age, B: Developmental noise parameter.

Table S5
Comparison to chance in each condition of Experiment 1, 2A and 2B

experiment	condition	mean	df	t_value	p_value
Experiment 1	test	0.74	39.00	5.51	$< .001$
Experiment 1	control	0.46	39.00	-0.94	$= .351$
Experiment 2A: Preference	test	0.97	39.00	29.14	$< .001$
Experiment 2A: Preference	control	0.64	39.00	2.70	$= .01$
Experiment 2B: Novelty	test	0.83	39.00	6.77	$< .001$
Experiment 2B: Novelty	control	0.59	39.00	1.49	$= .144$

Note. Proportion expected by chance = 0.5. Data aggregated within participant and conditon.

Finally, Figure S8 shows correlations between model predictions and the data, binned by year. Across noise parameters, model predictions and data are closest aligned (i.e. closest to the dotted line) for the pragmatics model, thereby corroborating the conclusions drawn based on the model comparison and the evaluation of the noise parameters. Correlations are also higher for 4yo compared to 3yo, supporting the interpretation based on the developmental noise parameter that children behaved more in line with the model predictions as they got older.

Supplemental results

The following sections present details on the analyses used to evaluate whether the different experimental manipulations produced different responses. For all generalized linear mixed models (GLMM) we used the maximally converging random effects structure.

Experiment 1

We used one sample t-tests to test if participants selected the more informative object (the one unique to the table the animal pointed at) above chance (50%). Table S5 shows the results.

To compare the two conditions, we fit the following GLMM: `correct ~ condition + (1 |id)`. Participants chose the more informative object more often in the test condition compared to the control condition ($\beta = 1.28$, se $= 0.29$, $p < .001$).

Experiment 2

Table S5 shows results for the comparison to chance in Experiment 2A and B. In Experiment 2A, we fit the following GLMM: `correct ~ condition + (1|id) + (1|agent)`. Here, `agent` refers to the animal making the request. Participants chose the object previously preferred by the speaker more often in the test (same speaker) condition compared to the control (different speaker) condition ($\beta = 2.92$, se $= 0.57$, $p < .001$). For Experiment 2B, we fit the following GLMM: `correct ~ condition + (condition|id)`. Participants chose the novel object more often in the test (same speaker) condition compared to the control (different speaker) condition ($\beta = 6.27$, se $= 1.96$, $p = .001$).

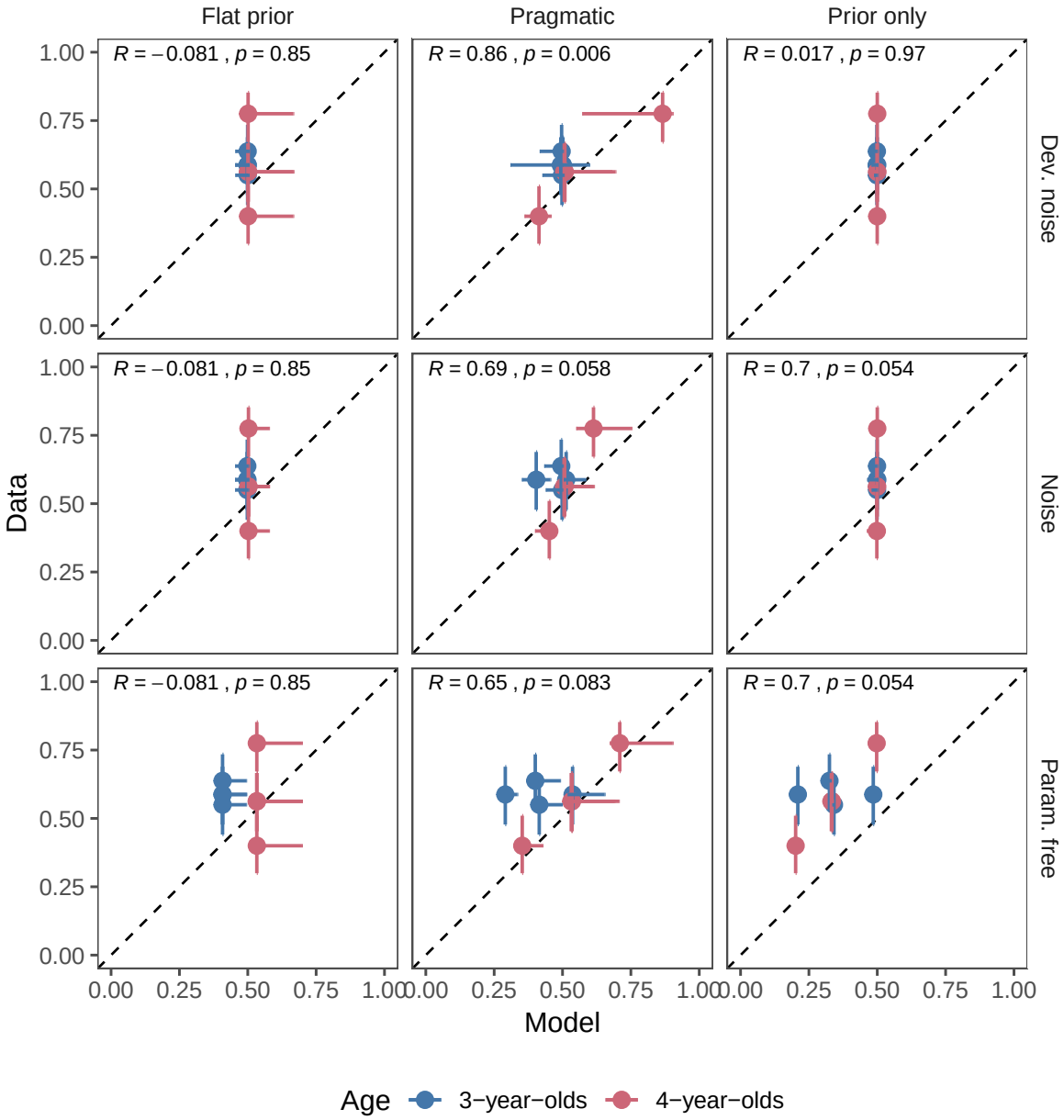

Figure S8. Correlation plot for model predictions and data for all models considered in Experiment 7. Dots represent condition modes. Error bars represent 95% HDIs.

Table S6
GLMM output for Experiment 3

term	estimate	std.error	statistic	p.value
Intercept	1.31	0.26	5.13	$< .001$
common ground	-0.50	0.34	-1.48	$= .14$
speaker	0.68	0.34	1.98	$= .048$
alignment	-0.31	0.39	-0.79	$= .427$
common ground * speaker	1.21	0.54	2.25	$= .025$
common ground * alignment	-0.91	0.53	-1.71	$= .088$
speaker * alignment	-2.64	0.48	-5.48	$< .001$
common ground * speaker * alignment	-1.05	0.72	-1.46	$= .145$

Note. Reference levels: commmon ground - novelty, speaker - different speaker, alignment - congruent.

Experiment 3

To see if participants differentiated between conditions, we fit the following GLMM to the data: `correct_inf ~ common ground*speaker*alignment + (alignment|id)`. The dependent variable `correct_inf` captures whether participants chose the more informative object. Table S6 shows the model output.

Prior strength manipualtions Experiment 4

Below we describe the different ways in which prior strength was manipulated in Experiment 4. The corresponding experiments can be found in the online repository. The test event was always the same: The animal disappeared and then either the same or a different animal returned and requested an object using an unknown word.

Table S7
GLMMs output for prior strength experiments

Manipulation	Strength	Term	Estimate	SE	p
novelty	medium	Intercept	0.36	0.29	$= .214$
novelty	medium	condition (same speaker)	0.67	0.31	$= .031$
novelty	strong	Intercept	0.51	0.32	$= .113$
novelty	strong	condition (same speaker)	1.59	0.40	$< .001$
novelty	weak	Intercept	0.10	0.26	$= .694$
novelty	weak	condition (same speaker)	0.16	0.29	$= .592$
preference	medium	Intercept	0.48	0.33	$= .145$
preference	medium	condition (same speaker)	1.74	0.40	$< .001$
preference	strong	Intercept	0.65	0.33	$= .046$
preference	strong	condition (same speaker)	3.61	0.79	$< .001$

Note. Model structure in all cases: correct ~ condition + (1|id) + (1|agent)

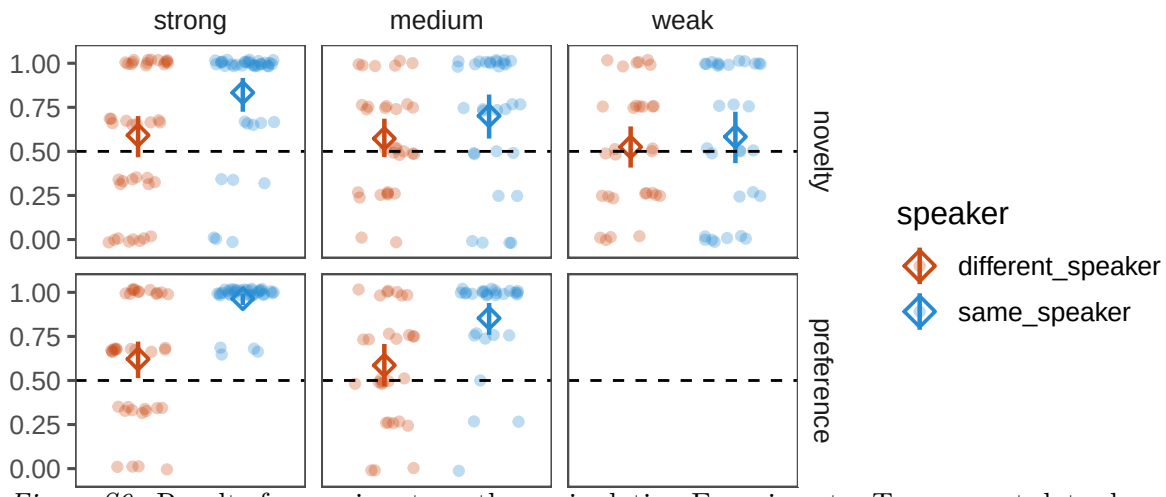

Figure S9. Results from prior strength manipulation Experiments. Transparent dots show data from individual participants, diamonds represent condition means, error bars are 95% CIs. Dashed line indicates performance expected by chance.

For preference, at the beginning of the experiment both tables contained an object. In preference/strong the animal turned to one side and stated that they liked ("Oh wow, I really like that one") or disliked ("Oh bleh, I really don't like that one") the object. Then they turned the other side and expressed the respective other attitude. In preference/medium the animal only turned to one side and expressed liking in a more subtle way (saying only: "Oh, wow").

For novelty, one table was empty while there was an object on the other. In novelty/strong the animal turned to one of the sides and commented either on the presence ("Aha, look at that") or the absence of an object ("Hm, nothing there"). Then the animal turned to the other side and commented in a complementary way. Next, the animal disappeared. The same animal re-appeared and the sequence above was repeated. When the animal disappeared for the second time, a second object appeared on the empty table while the animal was away. In novelty/medium, the animal commented on the presence/absence of objects in the same way but did so only once. In novelty/weak, the animal only turned to the present object and commented on it.

In all cases, the order of utterances and/or the side to which the speaker turned first were counterbalanced. Figure S9 shows the results for the same speaker and different speaker conditions for each manipulation.

Table S7 shows the results of a GLMM fit to the data from each manipulation. The results show that the parameter estimates for condition (i.e. difference between same speaker and different speaker condition) decreases in line with the hypothesized effect of the prior manipulation.

Experiment 4

To assess how participants differentiated between combinations of manipulations, we fit a separate GLMM to the data of each level of prior strength manipulation. Please note that we did not find a weak manipulation for preference. Model outputs are shown in Table

Table S8
GLMMs output for Experiment 4

prior strength	term	estimate	std.error	statistic	p.value
strong	Intercept	1.41	0.32	4.41	$< .001$
strong	common ground	-0.80	0.41	-1.97	$= .049$
strong	speaker	0.30	0.39	0.78	$= .438$
strong	alignment	-0.70	0.49	-1.44	$= .149$
strong	common ground * speaker	2.09	0.56	3.71	$< .001$
strong	common ground * alignment	0.44	0.65	0.67	$= .503$
strong	speaker * alignment	-1.60	0.52	-3.10	$= .002$
strong	common ground * speaker * alignment	-2.96	0.73	-4.03	$< .001$
medium	Intercept	1.47	0.31	4.73	$< .001$
medium	common ground	-0.10	0.42	-0.25	$= .801$
medium	speaker	0.24	0.35	0.70	$= .486$
medium	alignment	-0.42	0.43	-0.96	$= .335$
medium	common ground * speaker	1.48	0.58	2.57	$= .01$
medium	common ground * alignment	-1.26	0.59	-2.13	$= .033$
medium	speaker * alignment	-1.91	0.48	-3.99	$< .001$
medium	common ground * speaker * alignment	-1.36	0.74	-1.83	$= .068$
weak	Intercept	1.17	0.39	2.97	$= .003$
weak	speaker	-0.60	0.39	-1.54	$= .124$
weak	alignment	0.48	0.51	0.94	$= .345$
weak	speaker * alignment	-0.66	0.53	-1.24	$= .216$

Note. Reference levels: commmon ground - novelty, speaker - different speaker, alignment - congruent.

S8. Model structure for each level of prior strength manipulation was as follows:

- Strong:`correct_inf ~ common ground*speaker*alignment + (speaker+alignment|id)`
- Medium:`correct_inf ~ common ground*speaker*alignment + (alignment|id)`
- Weak:`correct_inf ~ correct_inf ~ speaker*alignment + (speaker+alignment|id)`

Experiment 5

Data for children was binned by year for the comparison to chance. As for adults, we used one sample t-tests for this analysis. Table S9 shows the results for Experiment 5.

We fit a GLMM with the following structure to the trial by trial data with age as a continuous variable: `correct ~ age_num + (1|id)`. Results show that children became more likely to select the more informative object with age ($\beta = 0.38$, se $= 0.11$, $p < .001$.

Table S9
Comparison to chance in each condition of Experiment 5 and 6

experiment	condition	mean	age_bin	df	t_value	p_value
Experiment 5	test	0.46	3.00	31.00	-1.31	= .198
Experiment 5	test	0.62	4.00	29.00	2.80	= .009
Experiment 6	control	0.47	3.00	29.00	-0.66	= .514
Experiment 6	control	0.50	4.00	30.00	0.00	=
Experiment 6	test	0.60	3.00	29.00	1.62	= .117
Experiment 6	test	0.71	4.00	30.00	4.14	< .001

Note. Proportion expected by chance = 0.5. Data aggregated within participant and conditon.

Table S10
GLMM output for Experiment 7

term	estimate	std.error	statistic	p.value
Intercept	0.46	0.19	2.34	= .019
age	-0.10	0.19	-0.50	= .615
speaker	0.50	0.28	1.79	= .073
alignment	-0.20	0.28	-0.73	= .468
age * speaker	0.57	0.27	2.08	= .037
age * alignment	0.12	0.28	0.45	= .655
speaker * alignment	-0.79	0.37	-2.15	= .032
age * speaker * alignment	-0.89	0.36	-2.47	= .013

Note. Reference level: speaker - different speaker, alignment - congruent.

Experiment 6

Comparison to chance within each condition of Experiment 6 is given in Table S9. We fit a GLMM with the following structure to the data: `correct ~ age_num*condition + (condition|id) + (1|agent)`. The model output shows an effect of condition (same or different speaker) (β = 0.89, se = 0.24, $p < .001$) but no effect of age (β = 0.02, se = 0.16, $p = .92$) or interaction between speaker identity and age (β = -0.01, se = 0.23, $p = .97$).

Experiment 7

To see if children differentiated between the different combinations of manipulations in Experiment 7, we fit the following GLMM to the data: `correct_inf ~ age_num*speaker*alignment+ (speaker+alignment|id)`. Model output is shown in Table S10.

References

Braginsky, M., Tessler, M. H., & Hawkins, R. (2019). *Rwebppl: R interface to webppl*. Retrieved from https://github.com/mhtess/rwebppl

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, *5*(1), 173–184.

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