Supplementary information:Predicting information integration in pragmatic word learning across development

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

Structure

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

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: the type of object they are (visually discernible given their shape and color) and their location (one of two tables). There are two types of objects t_1, t_2 and two locations l_1 and l_2. The utterances available to a speaker are (action, label) pairs, 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. The action of turning (pointing) reduces the set of referents to those that are on the table targeted by the turn (point); interpreting a label reduces that set further to those which are of the type referred to by that label. The interpretation of a label requires a lexicon $\mathcal{L}$, which provides the label–type mappings. There are two lexica: one which maps label 1 w_1 to type 1 t_1 and label 2 w_2 to type 2 t_2, and one 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. 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:

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

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 that a speaker would produce an utterance u given a referent r ($P_{S1}(u|r)$). 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 below for 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 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:

$$P_{S_1}(u|r) \propto exp(\alpha Informativity(u; r)) \tag{2}$$

As spelled out in (**???**), the informativity of an utterance for a referent is the (log) probability that a naive listener L_0 would select that referent given that utterance.

$$Informativity(u; r) = \ln P_{L_0}(r|u) \tag{3}$$

We assume that the speaker believes the literal listener to also be uncertain about the lexicon; that is, the pragmatic listener believes that the speaker (correctly) believes that the listener does not know the lexicon. Because the literal listener is assumed to not know the lexicon, the informativity of the utterance averages over the possible lexical, or words meanings, that L_0 considers:

$$P_{L_0}(r|u) = \sum_{L} P_{L_0}(r, \mathcal{L}|u) \tag{4}$$

where $P_{L_0}(r, \mathcal{L}|u) \propto \mathcal{L}_{lit}P(r)P(\mathcal{L})$. $P(r)$ again denotes the prior probability of a referent, $\mathcal{L}_{lit}$ encodes the literal meaning of an utterance given a particular lexicon, returning 1 if the utterance is literally true and 0 if the utterance is literally false. 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 (the turning or 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". That is, $\mathcal{L}_{lit}$ returns 1 for a referent that is on the table the agent turned to and 0 for referents on the other table.

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 the preference - different speaker - congruent

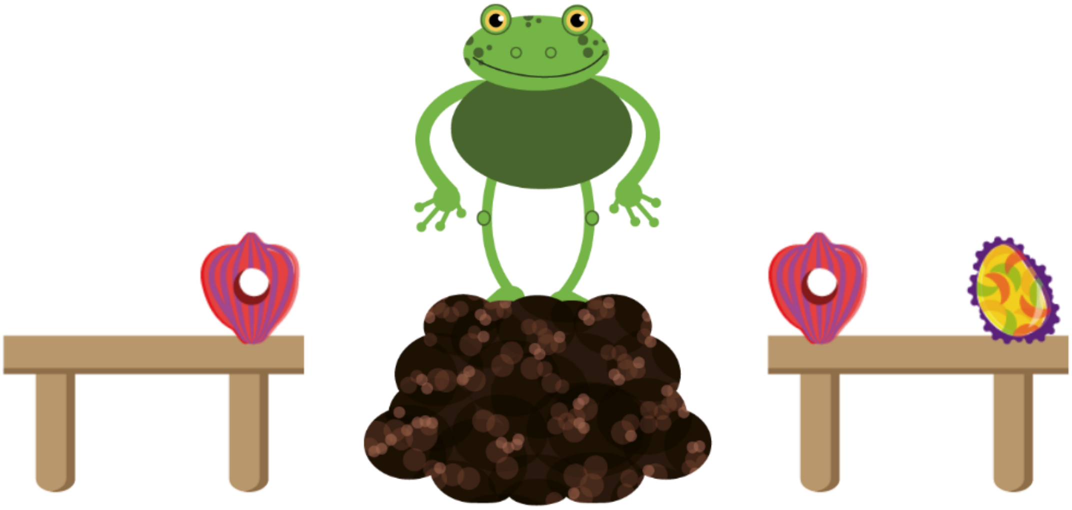

Figure 1. Screenshot from adult experiment.

condition (see below). Fig. 1 shows a screenshot from the adult experiment, which the model was designed to capture. In this context, there are three potential referents (two red-ish: r_1^r, r_2^r; one yellow-ish, r_3^y,), of two types (red-ish and yellow-ish) on two tables. In principle, the speaker, the frog in this case, can produce one of four utterances, turning 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, turning 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). <– CHECK ME.

The listener is learning the mappings between labels and object types (rather than object tokens). That is, the listener either thinks the novel word "dax" refers to "red-ish objects" or "yellow-ish objects". Turning or pointing to a table always has the same meaning. These semantics can be described using two lexica: $\mathcal{L}_\infty = \{dax : \text{red-ish thing}, wug : \text{yellow-ish thing}, point : \text{location of point}\}$, $\mathcal{L}_\in = \{dax : \text{yellow-ish thing}, wug : \text{red-ish thing}, point : \text{location of point}\}$.

We construct the prior distribution over referents $P(r)$ based on the results of Experiment 2A, in which the speaker displayed a preference for a yellow object (see Model Parameters section below for a detailed description of how this distribution was constructed). The prior distribution over referents $P(r)$ (left to right in Fig.1) for this condition was [0.26, 0.26, 0.48].

The literal listener's posterior distribution over referents and lexica has support of size six, owing to the unique combinations of referents and lexica.

$$\begin{aligned}
P_{L_0}(r_1^r, \mathcal{L}_1|u_3) &\propto \mathcal{L}_1(r_1^r, u_3)P(r_1^r)P(L_1) = 0 \times 0.26 \times 0.5\\
P_{L_0}(r_2^r, \mathcal{L}_1|u_3) &\propto \mathcal{L}_1(r_2^r, u_3)P(r_2^r)P(L_1) = 1 \times 0.26 \times 0.5\\
P_{L_0}(r_3^y, \mathcal{L}_1|u_3) &\propto \mathcal{L}_1(r_3^y, u_3)P(r_3^y)P(L_1) = 0 \times 0.48 \times 0.5\\
P_{L_0}(r_1^r, \mathcal{L}_2|u_3) &\propto \mathcal{L}_2(r_1^r, u_3)P(r_1^r)P(L_2) = 0 \times 0.26 \times 0.5\\
P_{L_0}(r_2^r, \mathcal{L}_2|u_3) &\propto \mathcal{L}_2(r_2^r, u_3)P(r_2^r)P(L_2) = 0 \times 0.26 \times 0.5\\
P_{L_0}(r_3^y, \mathcal{L}_2|u_3) &\propto \mathcal{L}_2(r_3^y, u_3)P(r_3^y)P(L_2) = 1 \times 0.48 \times 0.5
\end{aligned}$$

because the speaker is pointing to the table with r_2 and r_3, and under $\mathcal{L}_1$, “dax” means the red object and under $\mathcal{L}_2$, “dax” means the yellow object. After normalization we have:

$$\begin{aligned}
P_{L_0}(r_1^r, \mathcal{L}_1|u_3) &= 0\\
P_{L_0}(r_2^r, \mathcal{L}_1|u_3) &= 0.35\\
P_{L_0}(r_3^y, \mathcal{L}_1|u_3) &= 0\\
P_{L_0}(r_1^r, \mathcal{L}_2|u_3) &= 0\\
P_{L_0}(r_2^r, \mathcal{L}_2|u_3) &= 0\\
P_{L_0}(r_3^y, \mathcal{L}_2|u_3) &= 0.65
\end{aligned}$$

then,

$$\begin{aligned}
P_{L_0}(r_1|u_3) &= P_{L_0}(r_1, \mathcal{L}_1|u_3) + P_{L_0}(r_1, \mathcal{L}_2|u_3) = 0 + 0 = 0\\
P_{L_0}(r_2|u_3) &= P_{L_0}(r_2, \mathcal{L}_1|u_3) + P_{L_0}(r_2, \mathcal{L}_2|u_3) = 0.35 + 0 = 0.35\\
P_{L_0}(r_3|u_3) &= P_{L_0}(r_3, \mathcal{L}_1|u_3) + P_{L_0}(r_3, \mathcal{L}_2|u_3) = 0 + 0.65 = 0.65
\end{aligned}$$

and

$$\begin{aligned}
Informativity(u_3; r_1) &= log(0) = -\infty\\
Informativity(u_3; r_2) &= log(0.35) = -1.04\\
Informativity(u_3; r_3) &= log(0.65) = -0.43
\end{aligned}$$

Next we can insert these values into equation 2 to compute the probability that the speaker chooses this utterance to refer to this type of object. To do so, we need the speaker optimality parameter of $\alpha = 2.24$, which was estimated based on Experiment 1.

TO CREATE S(u | r), WE NEED TO HAVE THE INFORMATIVITY OF EACH OF THE 4 UTTERANCES. PASSING THE INFORMATIVITY THROUGH THE SOFT-MAX (speaker optimality) creates an “unnormalized” probability, which we then turn into a real probability by dividing by the sum of all the unnormalized probabilities of each utterance.

$$P_{S_1}(u|r) = \frac{exp(\alpha Informativity(u; r))}{\sum_{u'} exp(\alpha Informativity(u'; r))}$$

SEEMS LIKE WE NEED TO NORMALIZE ACROSS ALL UTTERANCES, SO WE WILL NEED TO REDO THE ABOVE CALCULATIONS FOR EACH OF THE 6 UTTERANCES.

$$P_{S_1}(u_3|r_2) = \frac{exp(2.24 \cdot -1.03)}{exp(2.24 \cdot -1.03) + exp(2.24 \cdot 0)} = 0.09$$
$$P_{S_1}(u_3|r_3) = \frac{exp(2.24 \cdot -0.44)}{exp(2.24 \cdot -0.44)} = 1$$

Here, the informativity for a referent is normalized across all utterances that could be used to refer to this referent.

Finally, based on equation 1 we can use these values to compute the probability that the speaker thinks that the speaker is referring to the yellow object (r_1) when they produce r_1:

$$P_{L_1}(r_1, L|u_1) = \frac{1 * 0.64}{1 \cdot 0.64 + 0.09 \cdot 0.36 + 0.09 \cdot 0.36} = 0.91$$

Thus, we normalize across all potential states that the utterance could refer to.

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 thinks S 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 RSA model (assuming equal prior probability for each object) 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 RSA 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).

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.

For adults, this worked in the following way: For example, in Experiment 2, for the preference/same speaker condition, when the speaker indicated that they liked object A and disliked object B, the average proportion with which participants chose object A was 0.97 and for object B it was 0.03 respectively. In Experiment 3, this measurement determined the prior distribution over objects in cases whenever the the same manipulation was used (preference/same speaker). When utterance and common ground information were aligned (i.e. object A was the more informative object), the distribution of objects was (A,B,B). The corresponding prior distribution was therefore (0.97, 0.03, 0.03). When information sources were dis-aligned (i.e. object B was the more informative one), the object distribution was (B,A,A) and the prior distribution was (0.03, 0.97, 0.97). 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.

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 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.

Prior only model

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

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

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 (e.g. 0.33,0.33,0.33). This was the case for children as well as adults. The speaker optimality parameter was set in the same way as in the pragmatics model.

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.

For preference, both tables initially 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

Table 1
Model 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)

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 2 shows the results for the same speaker and different speaker conditions for each manipulation.

Table 1 shows the results of a generalized linear mixed model (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.

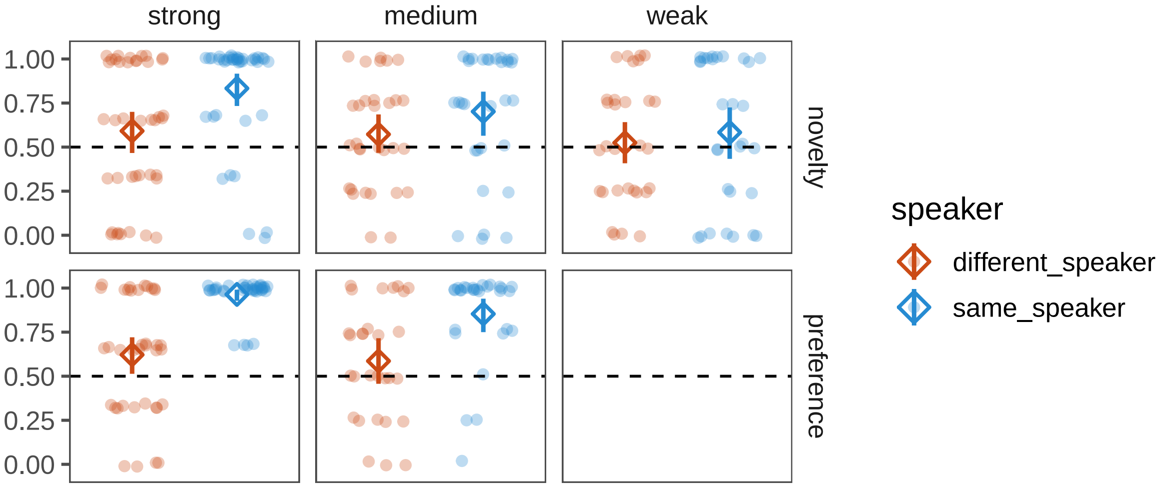

Figure 2. Results from prior strength manipulation Experiments. Transparent dots show data from individual participants, diamonds represent condition means, error bars are 95% CIs. Dashed line indicates performance expected by chance.

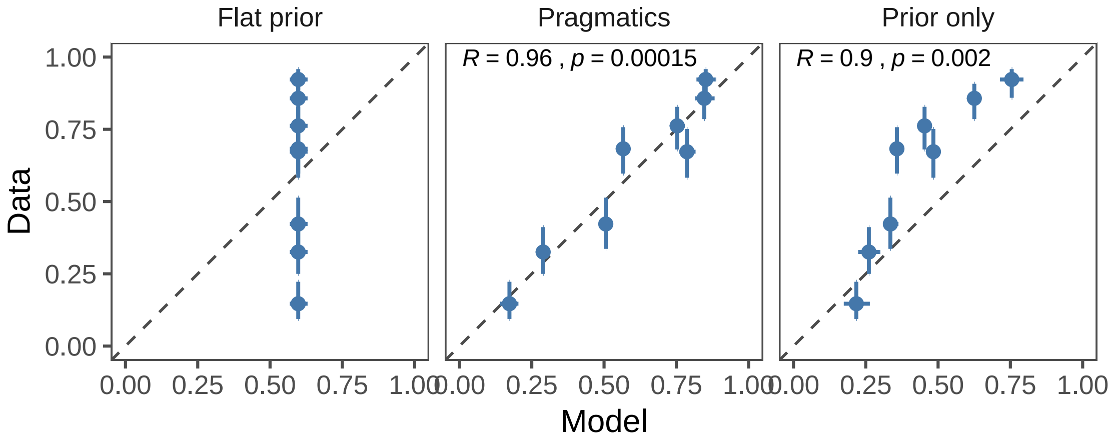

Figure 3. Correlation plot for model predictions and data from Experiment 3. All models depicted here included a noise parameter. Coefficients and p-values are based on Pearson correlation statistics. Dots represent condition modes. Error bars represent 95% HDIs.

Table 2
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

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 and then exponentiating the result. Table 2 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 2 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 3 compares model predictions from the models including noise parameters to the data from Experiment 3.

Figure 4 shows the posterior distribution of the noise parameter for 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

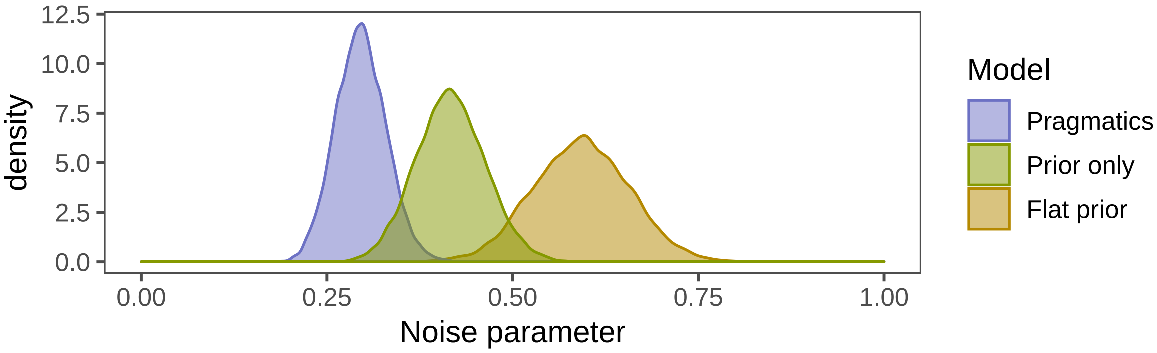

Figure 4. Posterior distribution of noise parameter for each model in Experiment 4

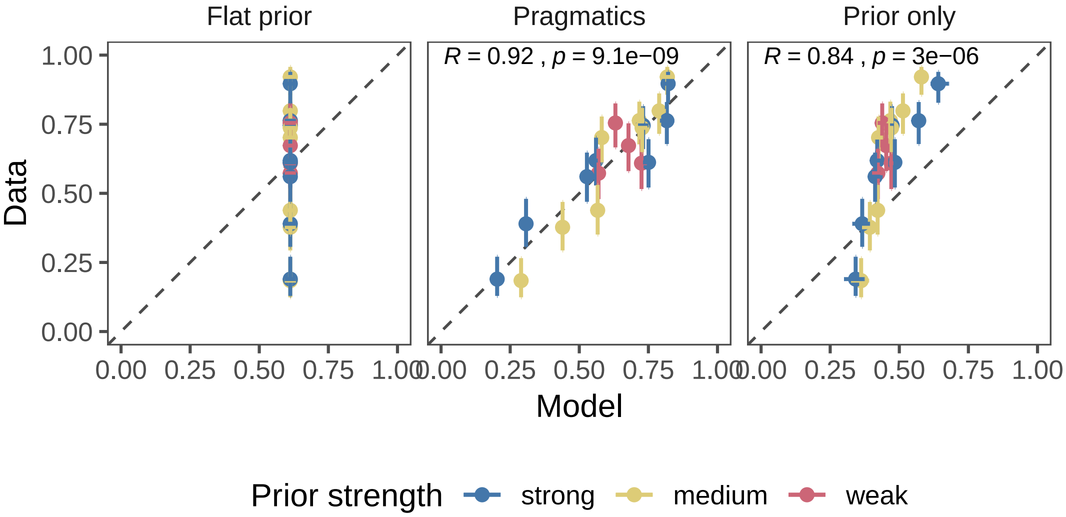

Figure 5. Correlation plot for model predictions and data from Experiment 4. All models included a noise parameter. Coefficients and p-values are based on Pearson correlation statistics. Dots represent condition modes. Error bars represent 95% HDIs.

lower noise parameter.

Experiment 4

Figure 5 compares model predictions to the data from Experiment 4. Table 3 shows Bayes Factors for model comparisons in Experiment 3. As pre-registered, all models included a noise parameter. Figure 6 shows the posterior distribution of the noise parameter for each model in Experiment 4.

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 comparably high. The additional model comparisons allow us to see if the pragmatics model provides a better fit

Table 3

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.

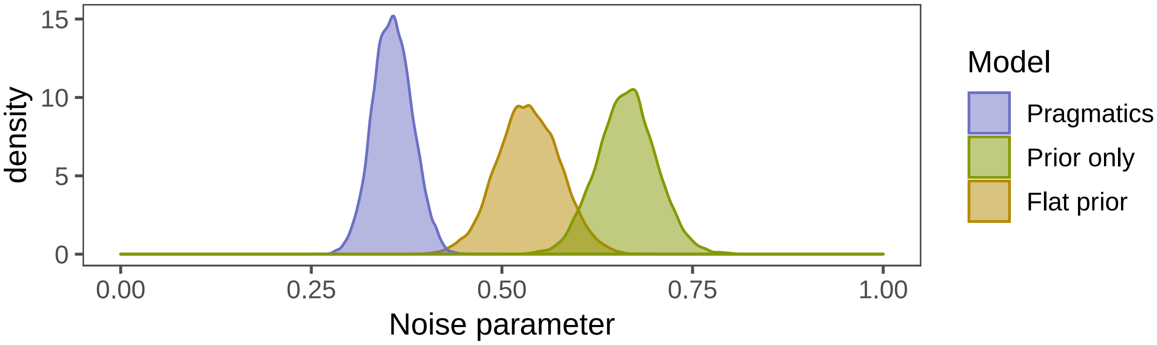

Figure 6. Posterior distribution of noise parameter for each model in Experiment 4

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 across age. 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 4 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 5 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

Table 4

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 5

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

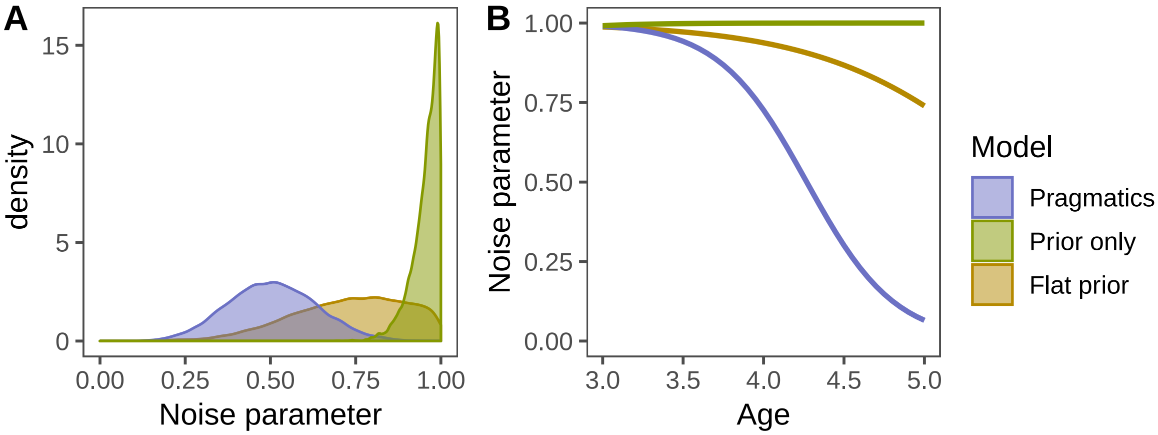

Figure 7. Posterior distribution of noise parameter for each model in Experiment 7. A: single noise parameter across age, B: Developmental noise parameter.

to the alternative models.

Figure 7 shows the different types of noise parameters for the each model. Figure 7A shows that the pragmatics model has the lowest estimated level of noise of all the models considered. Figure 7B 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.

Finally, Figure 8 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.

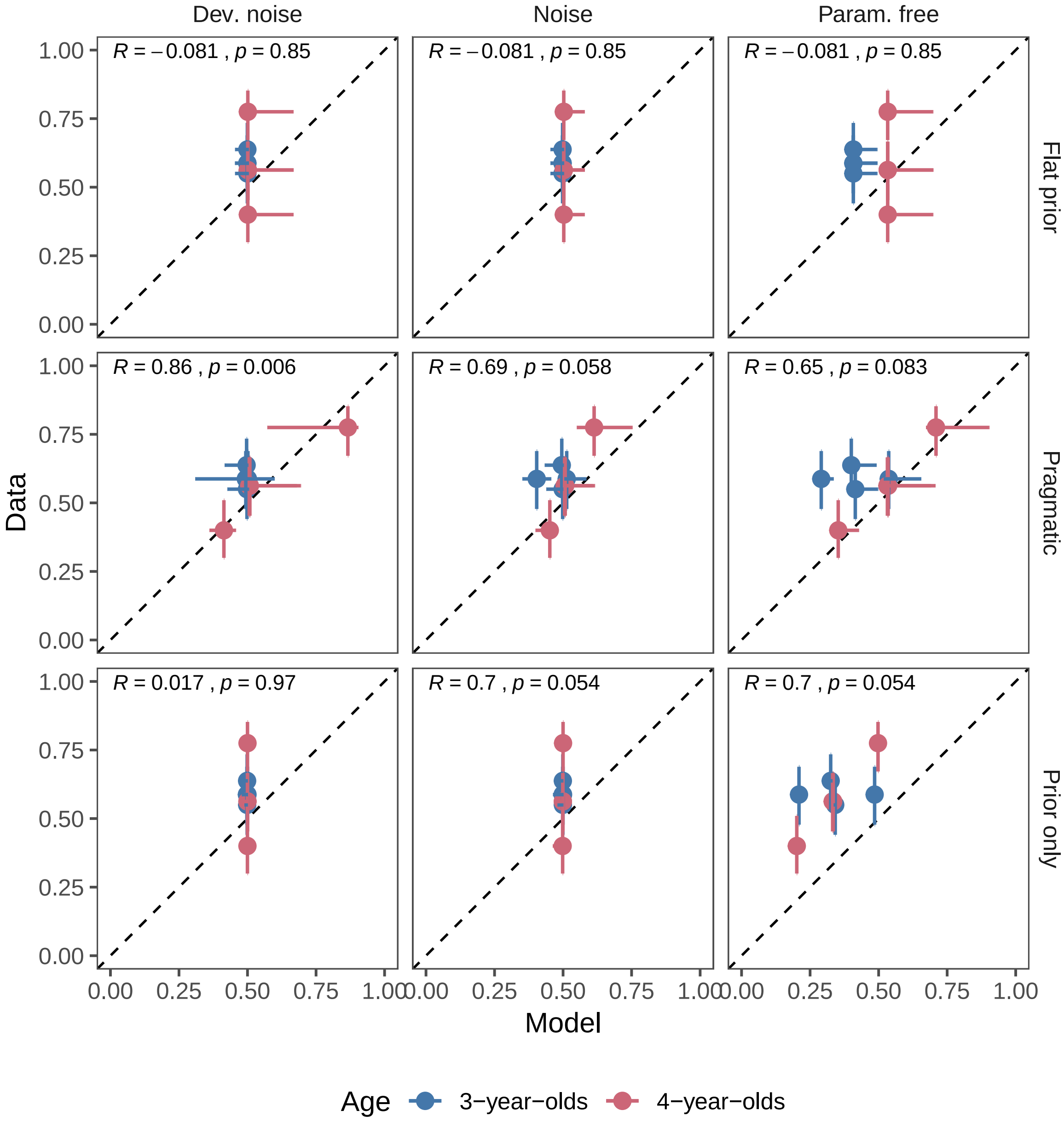

Figure 8. Correlation plot for model predictions and data for all models considered in Experiment 7. Dots represent condition modes. Error bars represent 95% HDIs.

References

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

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