Supplementary information
Predicting 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 additional information about methods, results and cognitive models. 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.

Supplemental methods and 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.

We used R (Version 3.6.1; R Core Team, 2018) and the R-packages *base* (Version 3.6.1; R Core Team, 2018), *broom* (Version 0.5.2; Robinson & Hayes, 2018), *coda* (Version 0.19.3; Plummer, Best, Cowles, & Vines, 2006), *dplyr* (Version 0.8.3; Wickham, François,

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

Henry, & Müller, 2019), *exactRankTests* (Version 0.8.30; Hothorn & Hornik, 2019), *forcats* (Version 0.4.0; Wickham, 2019a), *ggplot2* (Version 3.2.1; Wickham, 2016), *ggpubr* (Version 0.2.3; Kassambara, 2019), *ggthemes* (Version 4.2.0; Arnold, 2019), *knitr* (Version 1.25; Xie, 2015), *langcog* (Version 0.1.9001; Braginsky, Yurovsky, & Frank, n.d.), *lme4* (Version 1.1.21; Bates, Mächler, Bolker, & Walker, 2015; Bates et al., 2015), *lsr* (Version 0.5; Navarro, 2015), *magrittr* (Version 1.5; Bache & Wickham, 2014), *Matrix* (Version 1.2.17; Bates & Maechler, 2018), *matrixStats* (Version 0.55.0; Bengtsson, 2018), *papaja* (Version 0.1.0.9842; Aust & Barth, 2018), *purrr* (Version 0.3.2; Henry & Wickham, 2019), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *readxl* (Version 1.3.1; Wickham & Bryan, 2018), *rwebppl* (Version 0.1.997; Braginsky, Tessler, & Hawkins, 2019; Braginsky et al., 2019), *stringr* (Version 1.4.0; Wickham, 2019b), *tibble* (Version 2.1.3; Müller & Wickham, 2019), *tidyr* (Version 1.0.0; Wickham & Henry, 2019), and *tidyverse* (Version 1.2.1; Wickham, 2017) for all our analyses reported in the main manuscript and the supplementary material.

Experiment 1

Table S1

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.

Procedure. The setup and procedure of Experiment 1 for adults are shown in Figure 1A in the main manuscript as well as Figure S1. In the beginning of each trial, the animal introduced themselves (e.g. "Hi, I'm Dog") and then turned towards the table with the two objects. The same utterance was used to make a request in all adult studies ("Oh cool, there is a [non-word] on the table, how neat, can you give me the [non-word]?"). In the test condition, there was one object on the other table (left table in Figure S1), whereas in the control condition, there were two. In the control condition, no inference was possible based on the speaker's turning and pointing. The "correct" object in the control condition was randomly chosen from the two objects on the table. Technically, this condition did not control for any alternative explanations and we therefore did not run it for children. Participants received six trials, three per condition. Conditions were presented in a random order.

Results. 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 S1 and Figure S2 show 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 $= 0.29$, $p < .001$).
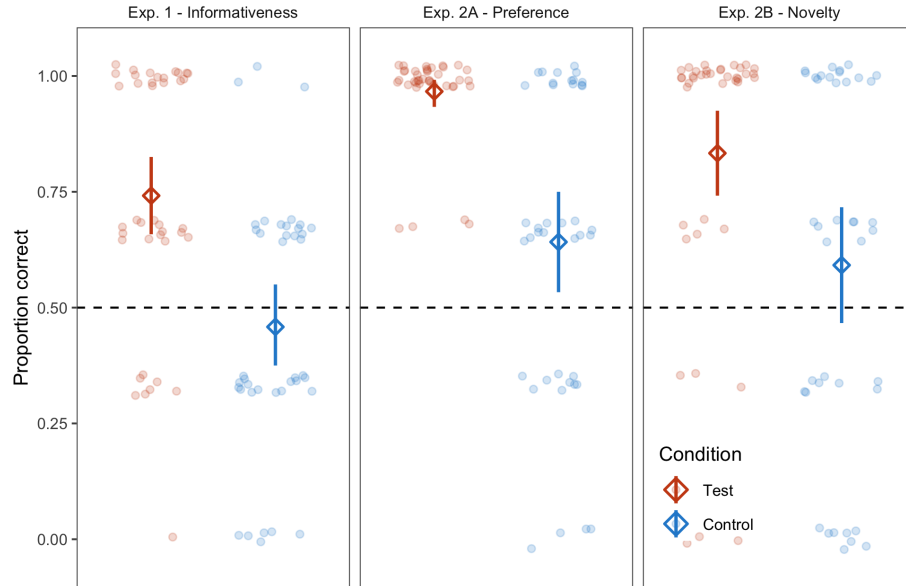
Figure S2. Results from experiment 1 and 2 for adults. For preference and novelty, control refers to a different speaker (see Fig. 1B). Transparent dots show data from individual participants, diamonds represent condition means, error bars are 95% CIs. Dashed line indicates performance expected by chance.

Experiment 2

Procedure.   The setup and procedure of Experiment 2 are shown in Fig. 1B in the main manuscript as well as Figure S3. In the preference manipulation, the animal introduced themselves, then turned to one of the tables and expressed either that they liked ("Oh wow, I really like that one") or disliked ("Oh bleh, I really don't like that one") the object before turning to the other side and expressing the respective other attitude. Next the animal disappeared and, after a short pause, either the same or a different animal returned and requested an object while facing straight ahead. Here, participants could identify the referent by assuming that the speaker was talking about the object they liked, if the speaker remained the same but not when a different speaker made a request. This procedure corresponds to the strong preference manipulation in Experiment 4 (see below).

In the novelty manipulation one of the tables was initially empty. The animal turned to one of the sides and commented either on the presence ("Aha, look at that") or the absence

Figure S3. Screenshot showing test situation in Experiment 2 for adults.

("Hm. . . , nothing there") of an object before turning to the other side and commenting in a complementary way. After shortly disappearing, the same animal repeated the sequence above. When the animal left a second time, a new object appeared on the empty table. Next, either the same or a different animal returned and requested an object. In this context, participants could infer that it was the new object that the speaker was referring to, if the speaker was the same. In contrast, when a different speaker (to whom both objects were equally new) made a request, no inference could be made based on the speaker's knowledge. This corresponded to the strong manipulation in Experiment 4 (see below).

In both Experiment 2A and 2B, participants always received six trials, three with the same and three with the different speaker. Conditions were presented in a random order.

Results. Table S1 and Figure S2 show 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 $= 1.96$, $p = .001$).

Experiment 3

   Procedure.   For Experiment 3 we inserted the common ground manipulation before the request in the setup of Experiment 1 (see Figure 1C in the main manuscript for one example). For example, the animal turned to the table with one object and expressed that they liked the object (pink-ish object in Figure S1), then turned to the other table and express that they did not like object B (yellow-ish object in Figure S1). Next, after quickly disappearing, the animal reappeared, turned to the table with two objects and made a request. To make it clear, which of the objects the speaker commented on while being turned to the table with the two objects during the common ground manipulation, the object was temporarily enlarged. Utterances were the same as in Experiment 1 and 2A and B.

   Combining information sources yielded eight different conditions. Informativeness and common ground could either be congruent or incongruent. If, for example, the yellow-ish object in Figure S1 (i.e. the more informative object because unique to that table) was also the one the speaker expressed preference for, the two sources of information were congruent. Next, the speaker who made the request could either be the same as the one in the previous interaction or a different one. Finally, common ground information could be manipulated in the form of preference or novelty. This resulted in a total of 2 (novelty or preference) x 2 (same or different speaker) x 2 (congruent or incongruent) = 8 conditions.

   Participants completed eight trials for one of the common ground manipulations with two trials per condition (same/different speaker x congruent/incongruent). Conditions were presented in a random order.

   Results.   We discuss and visualize the results as the proportion with which participants chose the more informative object (i.e., the object that would be the more informative referent when only utterance information is considered). FigureS4 shows the result in each

Table S2
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.
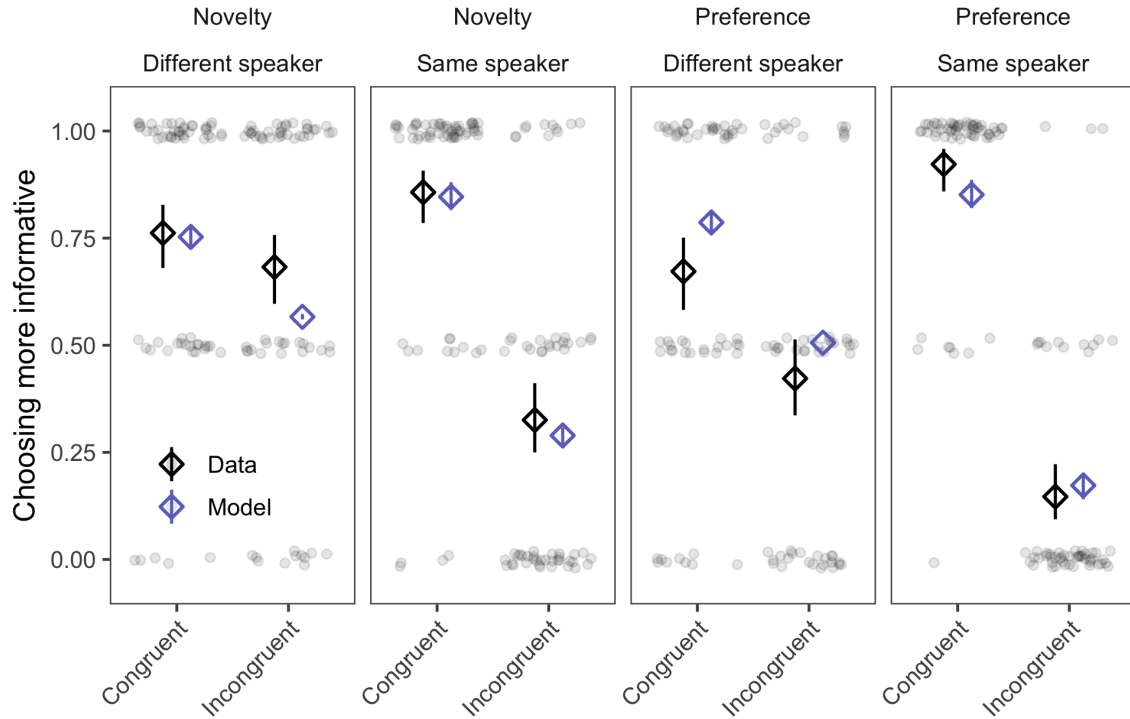

Figure S4. Data and model predictions by condition for Experiment 3 with adults. Transparent dots show data from individual participants, diamonds represent condition means. Empirical means are shonw in black and model predictions in pink. Error bars represent 95% HDIs.

Table S3
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)

of the eight conditions of Experiment 3. Results are plotted from The plot also shows model predictions from the pragmatics model (including the noise parameter - see below).

To see if participants differentiated between conditions, we fit the following GLMM to the data: `correct_inf ~ common ground*speaker*alignment + (alignment|id)`. Table S2 shows the model output. Results show that participants clearly differentiated between congruent and incongruent trials when the speaker remained the same.

Experiment 4

Prior strength manipualtions. 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, 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"). We piloted several procedures but did not find a manipualtion that indicated only weak preference.

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 S5 shows the results for the same speaker and different speaker conditions for each manipulation.

Procedure. The general procedure in Experiment 4 was the same as in Experiment 3. We ran a total of 2 (novelty or preference) x 3 (weak, medium, strong) x 2 (same or different speaker) x 2 (congruent or incongruent) - 4 (no weak manipulation for preference) = 20 conditions. Each participant was randomly assigned to a common ground manipulation and a level of prior strength and completed eight trials in total, two in each unique condition in that combination. The strong prior manipulation was a direct replication of Experiment 3.

Table S3 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.
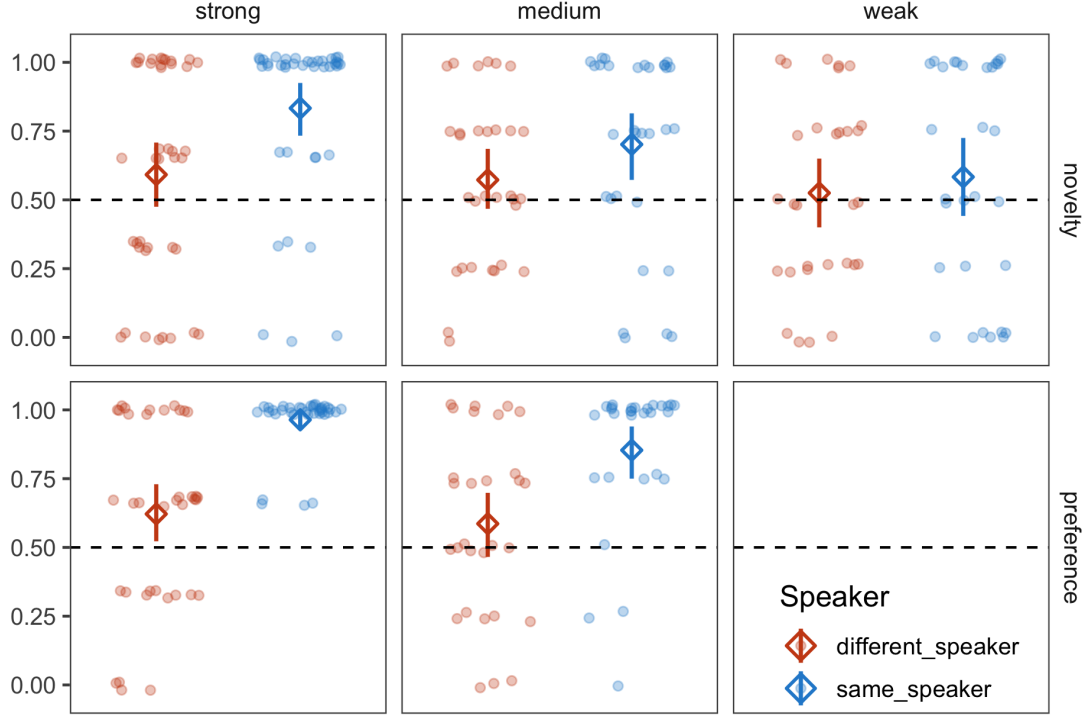
Figure S5. Results from prior strength manipulation Experiments. Transparent dots show data from individual participants, diamonds represent condition means, error bars are 95% CIs. Dashed line indicates performance expected by chance.

Results. Results from Experiment 3 and 4 (strong prior manipulation) were highly correlated (r = 0.97, $p > .001$).

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 S4. 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)`

Table S4

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.

A general observation to make is that, with common ground information getting weaker, changing the speaker has less of an impact on the informativeness inference. That is, the term `alignment x speaker` is only significant for the strong and medium manipulation but not the weak manipulation.

Experiment 5

Procedure. Experiment 5 for children was modeled after (M. C. Frank & Goodman, 2014). Instead of on tables, objects were presented as hanging in trees in order to facilitate showing points to distinct locations (see Figure S6). After introducing themselves, the animal turned to the tree with two objects and said: "This is a tree with a [non-word], how neat, a tree with a [non-word]"). Next, the trees and the objects in them disappeared and
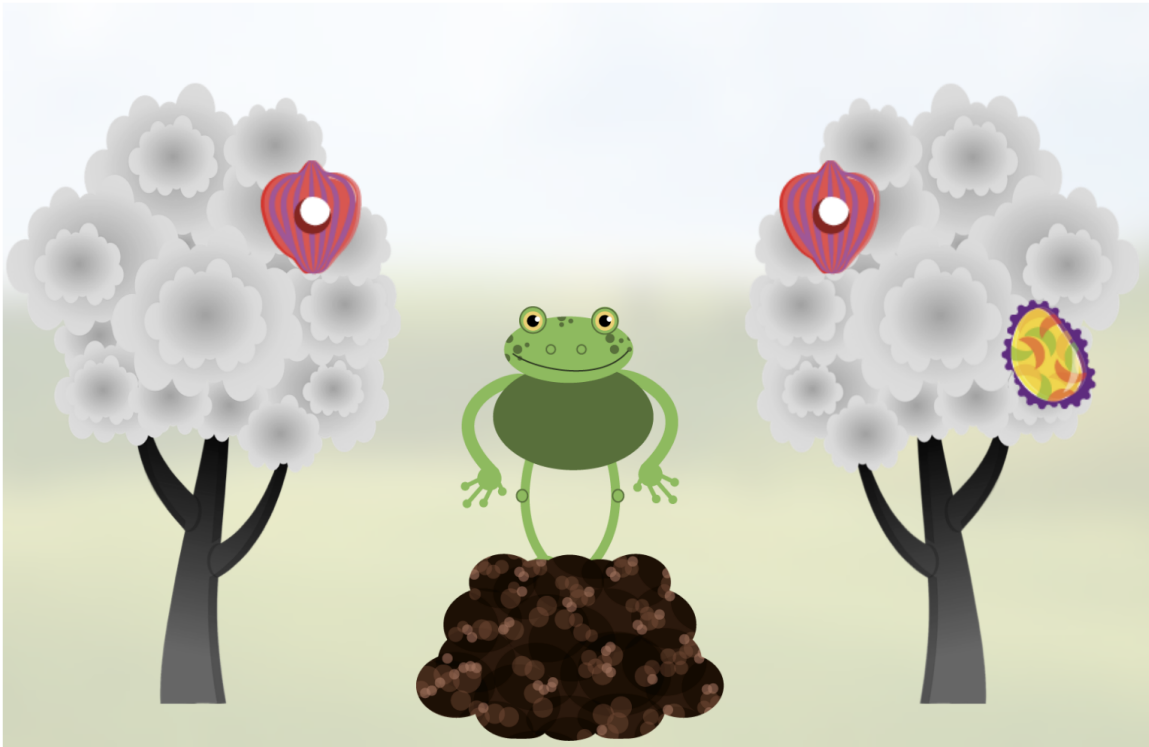
Figure S6. Screenshot showing test situation Experiment 5 and 7.

new trees replaced them. The two objects from the tree the animal turned to previously were now spread across the two trees (one object per tree, position counterbalanced). While facing straight, the animal first said "Here are some more trees" and then asked the child to pick the tree with the object that corresponded to the novel word ("Which of these trees has a [non-word]?"). Children received six trials in a single test condition.

Results. 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 S5 shows the results for Experiment 5. Figure S7 (left) shows the results plotted with age as a continuous variable.

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 $= 0.11$, $p < .001$, see also Figure S7).

Table S5

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

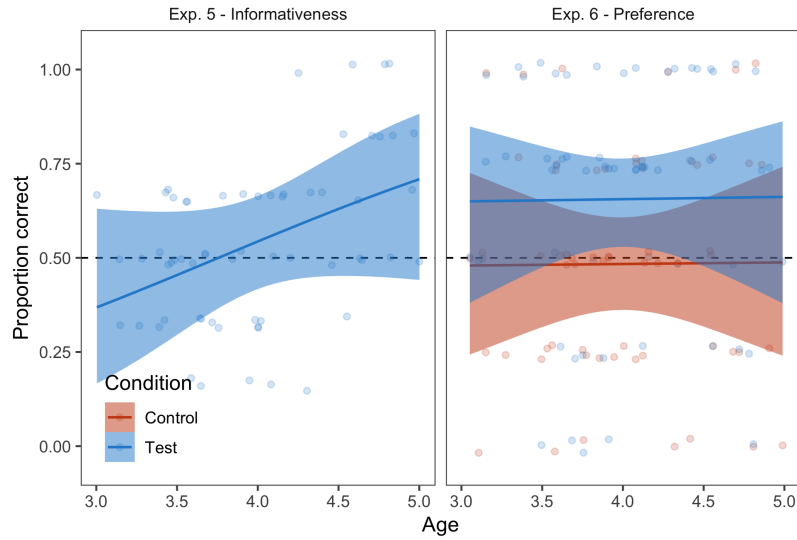*Note.* Proportion expected by chance = 0.5. Data aggregated within participant and conditon.

Figure S7. Results from experiment 5 and 6 for children. For preference, control refers to to the different speaker condition (see Fig. 1B in main manuscript). Transparent dots show data from individual participants, regression lines shows smoothed conditional mean with 95% CI. Dashed line indicates performance expected by chance.

Experiment 6

Procedure. Experiment 5 for children was identical to the strong preference manipulation for adults. Children received eight trials, four with the same and four with a different animal returning, presented in a random order. We tested children in the novelty manipulation but found that they did not reliably distinguish between the same and different speaker conditions.

Results. Comparison to chance within each condition of Experiment 6 is given in Table S5. 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) ($\beta = 0.89$, se $= 0.24$, $p < .001$) but no effect of age ($\beta = 0.02$, se $= 0.16$, $p = .92$) or interaction between speaker identity and age ($\beta = -0.01$, se $= 0.23$, $p = .97$, see also Figure S7).

Experiment 7

Procedure. In Experiment 7 for children, we again inserted the preference manipulation from Experiment 6 into the setup of Experiment 5. After greeting the child, the animal turned to one of the trees, pointed to an object (object was temporarily enlarged and moved closer to the animal) and expressed liking or disliking. Then the animal turned to the other tree and expressed the other attitude for the other kind of object. Next, the animal disappeared and either the same or a different animal returned. The rest of the trial was identical to the label and request phase of Experiment 5. Children received eight trials, two per condition (same/different speaker x congruent/incongruent) in a randomized order. Children had the option to opt out after half of the test trials. Therefore, we constraint the randomization of conditions so that children received two blocks of trials, one with a trial from each condition.

Table S6

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.
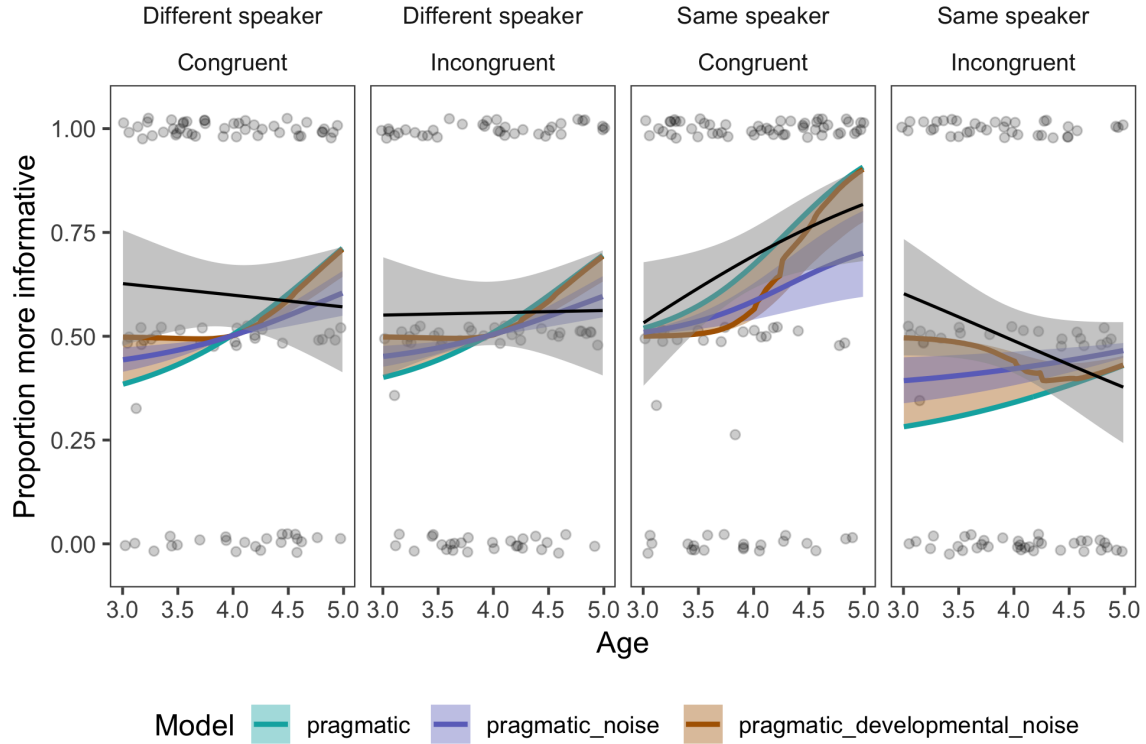

Figure S8. Model predictions and data for Experiment 7 with children. Colored lines show model predictions with 95% HDIs, transparent dots show data from individual participants and the black line with grey CI shows the smoothed conditional mean of the data in each condition.

Results.   Figure S8 shows the results for each condition plotted by age as a continuous variable. The same figure also shows model predictions from the pragmatics model with different noise parameters (see below).

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 S6.

Cognitive Models

Cognitive models were implemented in WebPPL (Goodman & Stuhlmüller, 2014) using the r package `rwebppl` (Braginsky et al., 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 (M. C. 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 $\sum_{r_i \in t} P_{L_0}(r_i|u)$.

$P(r)$ 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). Figure 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.

$$P_{L_0}(r_{p1}|u_3) \propto \mathcal{L}_{point}(r_{p1}, u_3)P(r_{p1}) = 0 \times 0.26 = 0$$

$$P_{L_0}(r_{p2}|u_3) \propto \mathcal{L}_{point}(r_{p2}, u_3)P(r_{p2}) = 1 \times 0.26 = 0.26$$

$$P_{L_0}(r_{y2}|u_3) \propto \mathcal{L}_{point}(r_{y2}, u_3)P(r_{y2}) = 1 \times 0.48 = 0.48$$

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

$$P_{L_0}(r_{p1}|u_3) = 0$$

$$P_{L_0}(r_{p2}|u_3) = 0.35$$

$$P_{L_0}(r_{y2}|u_3) = 0.65$$

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

$$Informativity(u_3; r_p) = P_{L_0}(r_p|u_3) = P_{L_0}(r_{p1}|u_3) + P_{L_0}(r_{p2}|u_3) = 0 + 0.35 = 0.35$$

$$Informativity(u_3; r_p) = P_{L_0}(r_y|u_3) = P_{L_0}(r_{y2}|u_3) = 0.65$$

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:

$$Informativity(u_1; r_p) = P_{L_0}(r_p|u_1) = 1$$

$$Informativity(u_1; r_y) = P_{L_0}(r_y|u_1) = 0$$

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.

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

$$P_{L_1}(r_{p1}, \mathcal{L}_1|u_3) \propto P_{S_1}(u_3 \mid r_{p1}, \mathcal{L}_1)P(r_{p1})P(\mathcal{L}_1) = 0.09 \times 0.26 \times 0.5 = 0.012$$

$$P_{L_1}(r_{p2}, \mathcal{L}_1|u_3) \propto P_{S_1}(u_3 \mid r_{p2}, \mathcal{L}_1)P(r_{p2})P(\mathcal{L}_1) = 0.09 \times 0.26 \times 0.5 = 0.012$$

$$P_{L_1}(r_{y2}, \mathcal{L}_1|u_3) \propto P_{S_1}(u_3 \mid r_{y2}, \mathcal{L}_1)P(r_{y2})P(\mathcal{L}_1) = 0 \times 0.48 \times 0.5 = 0$$

$$P_{L_1}(r_{p1}, \mathcal{L}_2|u_3) \propto P_{S_1}(u_3 \mid r_{p1}, \mathcal{L}_2)P(r_{p1})P(\mathcal{L}_2) = 0 \times 0.26 \times 0.5 = 0$$

$$P_{L_1}(r_{p2}, \mathcal{L}_2|u_3) \propto P_{S_1}(u_3 \mid r_{p2}, \mathcal{L}_2)P(r_{p2})P(\mathcal{L}_2) = 0 \times 0.26 \times 0.5 = 0$$

$$P_{L_1}(r_{y2}, \mathcal{L}_2|u_3) \propto P_{S_1}(u_3 \mid r_{y2}, \mathcal{L}_2)P(r_{y2})P(\mathcal{L}_2) = 1 \times 0.48 \times 0.5 = 0.24$$

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

$$P_{L_1}(r_{p1}, \mathcal{L}_1|u_3) = 0.045$$

$$P_{L_1}(r_{p2}, \mathcal{L}_1|u_3) = 0.045$$

$$P_{L_1}(r_{y2}, \mathcal{L}_1|u_3) = 0$$

$$P_{L_1}(r_{p1}, \mathcal{L}_2|u_3) = 0$$

$$P_{L_1}(r_{p2}, \mathcal{L}_2|u_3) = 0$$

$$P_{L_1}(r_{y2}, \mathcal{L}_2|u_3) = 0.91$$

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). 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. S3 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. S3) and disliked the pink-ish object (left table in Fig.S3), 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 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
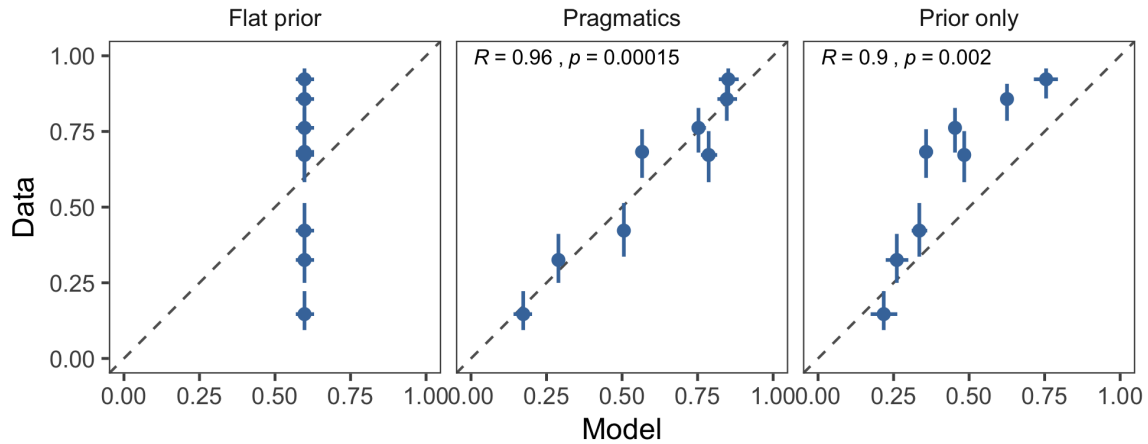
Figure S9. Correlation plot for model predictions and data from Experiment 3. All models depicted here included a noise parameter. Coefficients and p-values are based on Pearson correlation statistics. Dots represent condition modes. Error bars represent 95% HDIs.

first subtracting log-likelihoods for two models and then exponentiating the result. Table S7 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 S7 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 S9 correlates model predictions from the models including noise parameters to the data from Experiment 3.

Figure S10 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.

Table S7

Bayes Factors for model comparisons in Experiment 3

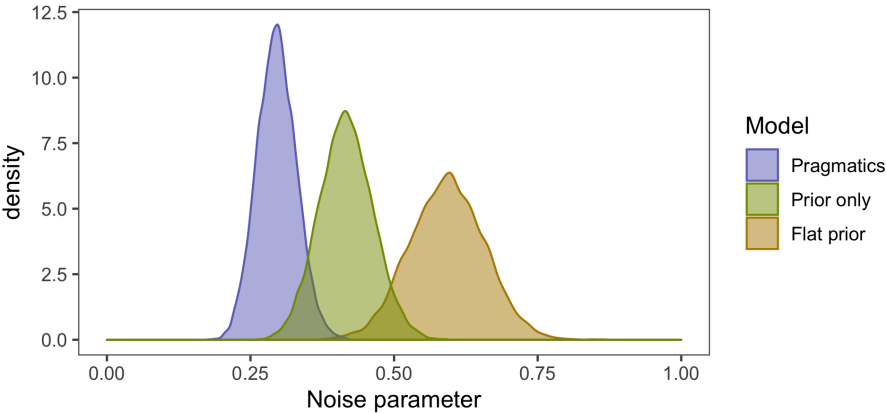| Comparison	BF
pragmatic_noise > pragmatic	2.6e+227
pragmatic_noise > prior_only_noise	2.5e+34
pragmatic_noise > flat_prior_noise	4.2e+53
prior_only_noise > flat_prior_noise	1.7e+19

Figure S10. Posterior distribution of noise parameter for each model in Experiment 4

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 S8

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

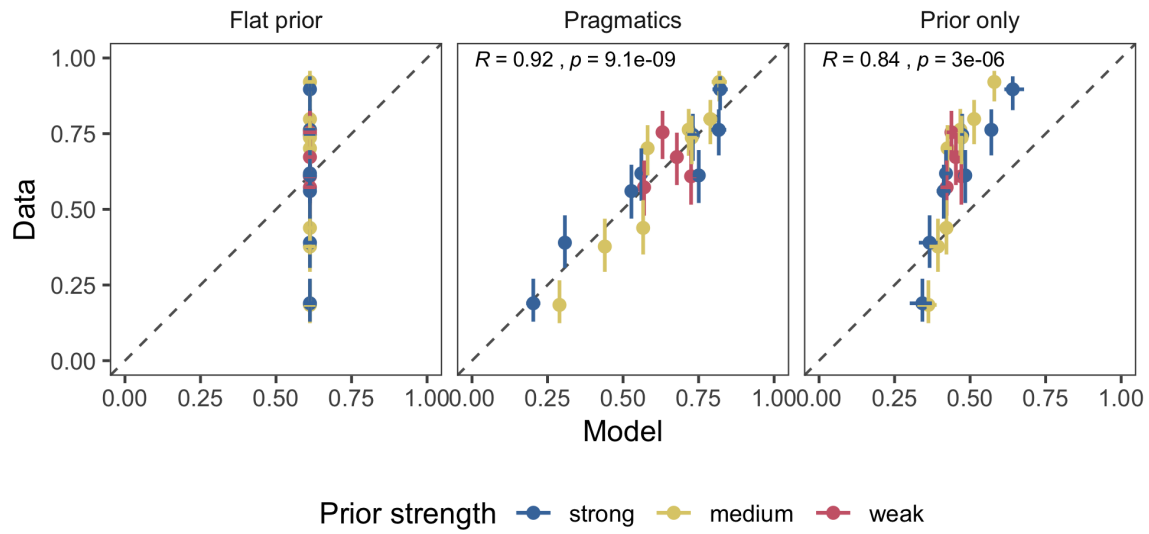*Note.* BF = Bayes Factor; All models include a noise parameter.

Figure S11. 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.
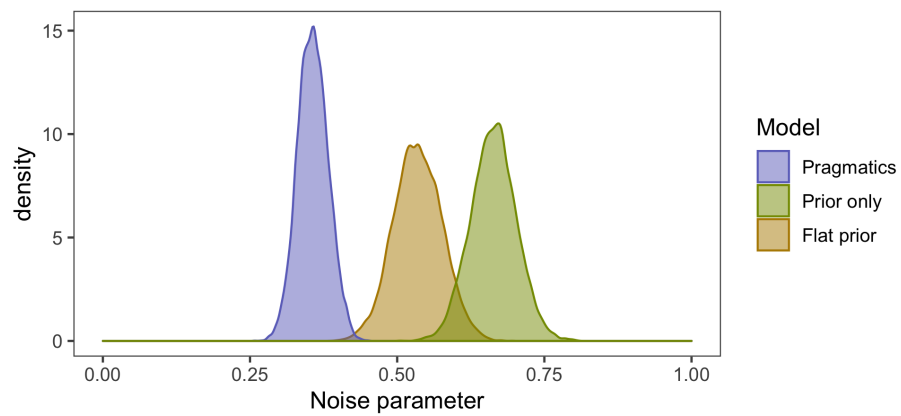

Figure S12. Posterior distribution of noise parameter for each model in Experiment 4

Experiment 4

Figure S11 correlates model predictions with the data from Experiment 4. Table S8 shows Bayes Factors for model comparisons in Experiment 4. As pre-registered, all models included a noise parameter. Figure S12 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 (see Figure S8 for model predictions from the models with different noise parameters). 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 S9 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 S10 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 S13 shows the different types of noise parameters for the each model. Figure S13A shows that the pragmatics model has the lowest estimated level of noise of all the

Table S9
*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 S10
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

models considered. Figure S13B 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 S14 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.

Note on alternative approach

In this model, we assume that the literal listener (L_0) is also sensitive to the prior manipulation. An alternative approach (used e.g. in M. C. Frank & Goodman, 2012) is to
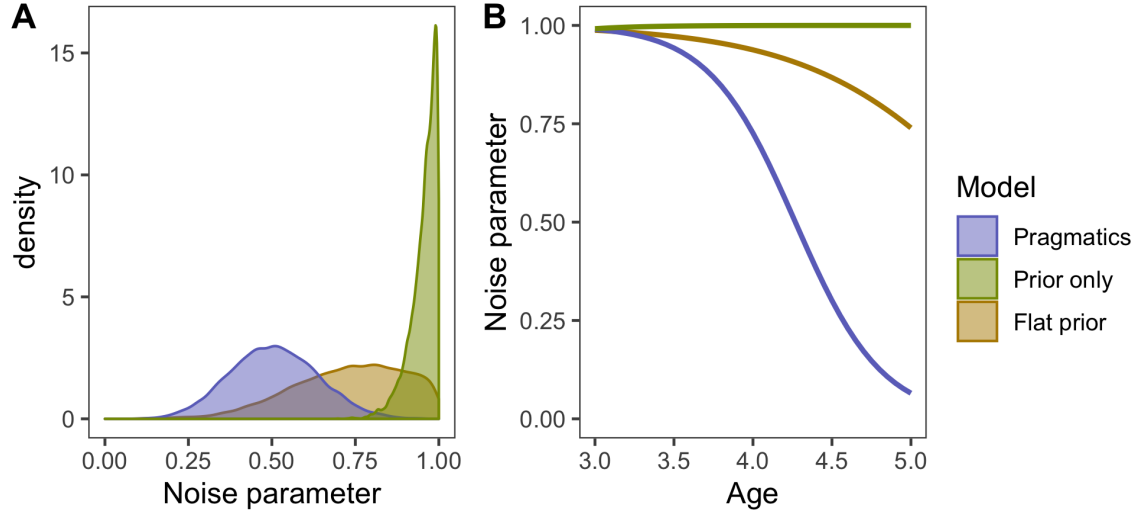
Figure S13. Posterior distribution of noise parameter for each model in Experiment 7. A: single noise parameter across age, B: Developmental noise parameter.

assume that L_0 has a uniform prior distribution over referents [.33, .33, .33]. We also ran the entire analysis with this alternative approach and found no major changes in the results. In all cases, the pragmatics model outperformed the two alternative models, both for children as well as for adults. Interestingly, for adults but not for children, using a uniform prior for L_0 yielded a lower estimate for the noise parameter in Experiment 3 and 4.
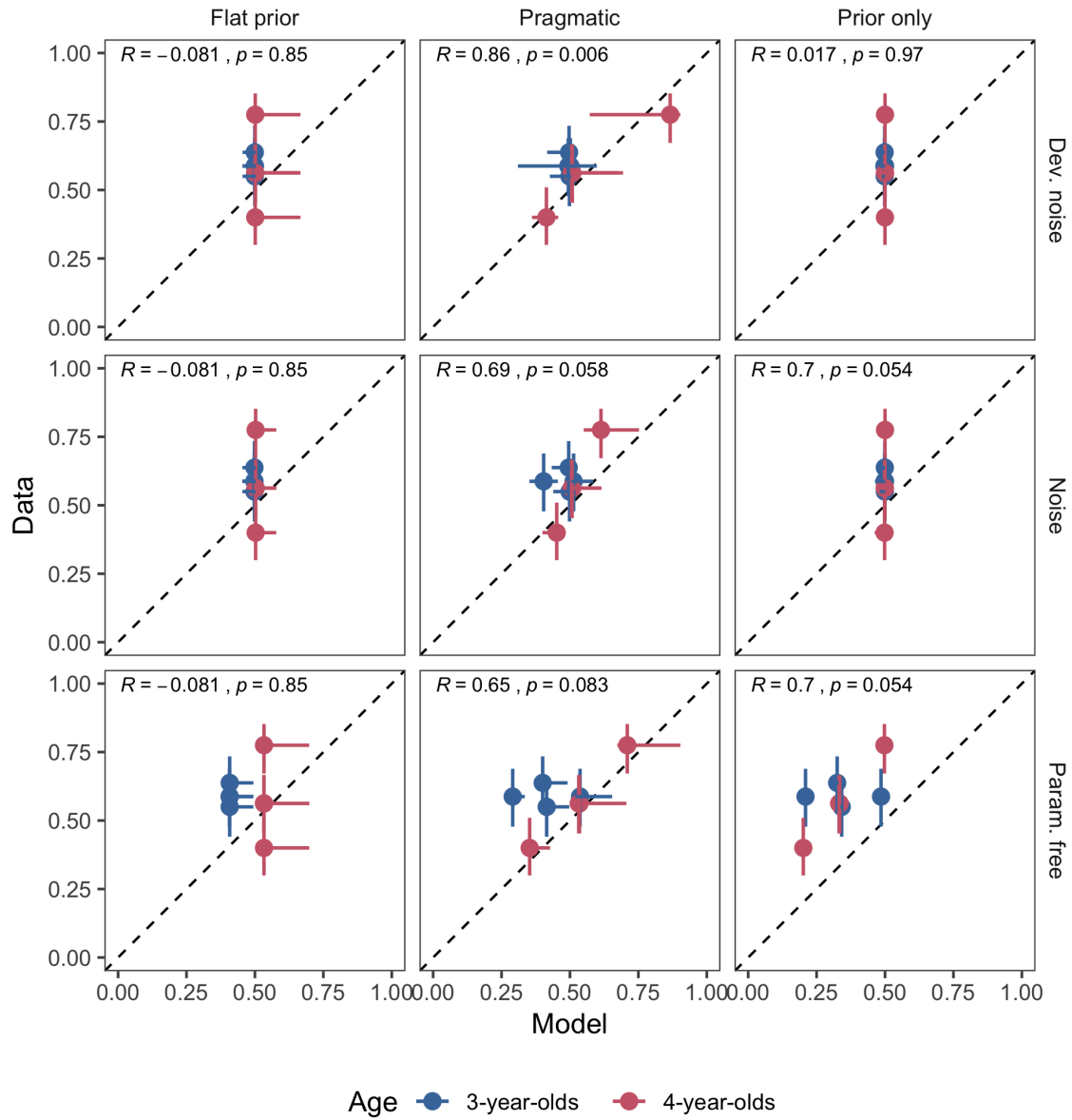
Figure S14. Correlation plot for model predictions and data for all models considered in Experiment 7. Dots represent condition modes. Error bars represent 95% HDIs.

References

Arnold, J. B. (2019). *Ggthemes: Extra themes, scales and geoms for 'ggplot2'.* Retrieved from https://CRAN.R-project.org/package=ggthemes

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown.* Retrieved from https://github.com/crsh/papaja

Bache, S. M., & Wickham, H. (2014). *Magrittr: A forward-pipe operator for r.* Retrieved from https://CRAN.R-project.org/package=magrittr

Bates, D., & Maechler, M. (2018). *Matrix: Sparse and dense matrix classes and methods.* Retrieved from https://CRAN.R-project.org/package=Matrix

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Bengtsson, H. (2018). *MatrixStats: Functions that apply to rows and columns of matrices (and to vectors).* Retrieved from https://CRAN.R-project.org/package=matrixStats

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

Braginsky, M., Yurovsky, D., & Frank, M. (n.d.). *Langcog: Language and cognition lab things.* Retrieved from http://github.com/langcog/langcog

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

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

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.

Henry, L., & Wickham, H. (2019). *Purrr: Functional programming tools.* Retrieved from https://CRAN.R-project.org/package=purrr

Hothorn, T., & Hornik, K. (2019). *ExactRankTests: Exact distributions for rank and permutation tests.* Retrieved from https://CRAN.R-project.org/package=exactRankTests

Kassambara, A. (2019). *Ggpubr: 'Ggplot2' based publication ready plots.* Retrieved from https://CRAN.R-project.org/package=ggpubr

Müller, K., & Wickham, H. (2019). *Tibble: Simple data frames.* Retrieved from https://CRAN.R-project.org/package=tibble

Navarro, D. (2015). *Learning statistics with r: A tutorial for psychology students and other beginners. (version 0.5).* Adelaide, Australia: University of Adelaide. Retrieved from http://ua.edu.au/ccs/teaching/lsr

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Robinson, D., & Hayes, A. (2018). *Broom: Convert statistical analysis objects into tidy tibbles.* Retrieved from https://CRAN.R-project.org/package=broom

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'.* Retrieved from

https://CRAN.R-project.org/package=tidyverse

Wickham, H. (2019a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2019b). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H., & Bryan, J. (2018). *Readxl: Read excel files*. Retrieved from https://CRAN.R-project.org/package=readxl

Wickham, H., & Henry, L. (2019). *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., François, R., Henry, L., & Müller, K. (2019). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., Hester, J., & Francois, R. (2018). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/