Communicative pressure on caregivers leads to language input that supports children's word learning

Ben Morris and Daniel Yurovsky

S1 Corpus Referential Talk and Gesture Coding Supplemental

Using the session transcripts, we coded all the concrete referents (in either speech or gesture) throughout the session that were produced by the parent or child. We further coded whether a referent was likely to be physically present, based primarily on transcript context. Coding was led by Madeline Meyers. A full coding manual can be found on the Open Science Foundation project page: https://osf.io/ums45/.

S1.1 Coding A Referent

In our coding, we aimed to capture any referential communication about concrete in gesture or speech. A concrete noun can roughly be thought of as any noun that could be pointed to if it was present. Note that a single utterance may have multiple referents. For gesture, any deictic gesture toward an identifiable referent definitionally refers to a concrete noun. For references in spoken language, note that the referent sometimes differs from the word in the utterance– for example, using a nickname like "bidiba" to consistently refer to a blanket would be coded as "blanket."

Conversely, non-concrete nouns can roughly be thought of as any noun that cannot be pointed to, such as abstract concepts. These nouns were not included. Gestures that could not be readily identified or that gestured to vague locations (e.g., "over there" while gesturing) also did not receive a referent code.

S1.2 Coding for Presentness

Not all concrete referents that are mentioned in speech are actually present in the environment. Using contextual information in the transcripts, we thus also coded whether each referent was likely to be present or not. Key contextual features included whether the same referent had been recently gestured to, question forms (e.g., "Where is the X?"), and semantic information about presentness ("Look at that?"). Every referent in an utterance is coded as "1" if present or "0" if not present. A subset of these codes were also done using video data to establish the reliability of our presentness coding. Further information can be found in the full coding manual on the Open Science Foundation project page: https://osf.io/ums45/.

S1.3 Coding Example

person	chat	spoken obj	ref_pres_predicted
parent	want a bite of banana ?	banana	1
child	orange .	orange	0
parent	that's all the orange we have .	orange	0
parent	that's it .		

S2 Experimental Results

For readability, the main text includes only the key effects for each statistical model rather than a full specification. We include those here. Each model included at least a random intercept for each subject and item.

S2.1 Learning

To confirm that we successfully manipulated participants' learning, we asked whether items with more exposure during training were better learned at pretest. To do this we fit a logistics mixed-effects model to analyze learning at baseline (i.e. prior to gameplay). We see the predicted significant effect of exposure rate on learning, confirming that object-label mappings that were presented more in training were better learned.

Additionally, we tested our critical between-subjects manipulations to ensure that learning of the lexicon did not differ significantly at pretest (prior to the manipulations). Neither utility condition or partner's exposure significantly predicted performance at pretest. This provides a simple check that participants in each condition learned the lexicon similarly. The full results of this model are presented in Table S1.

Table S1: Pariticpant learning at baseline, specified as `testCorrect` $\sim$ `exposureRate + condition + (exposureRate | subj) + (1 | realLabel)`.

term	estimate	std.error	z-value	p-value
intercept	-0.97	0.23	-4.21	$< .001$
exposure rate	1.08	0.08	13.67	$< .001$
utility condition	-0.22	0.19	-1.15	.249
partner's exposure	0.11	0.12	0.93	.352

S2.2 Communicative Strategy

Our key analyses concerned participants choice of communicative strategy. In each trial, participants were able to choose one of 3 communicative strategies: point, speak, or teach We expected flexible trade-off between the use of each strategy given their relative utilities, participants' knowledge of the lexicon, and the partners' knowledge of the lexicon. To test our predictions about each communicative behavior (gesture, speech, and teaching), we conducted separate logistic mixed-effects models for each behavior, reported below. The mixed effects model for each communicative

behavior has an identical effect structure for comparability. Each model included a random effect of subject and item. It should be noted that these three behaviors are mutually exhaustive.

S2.2.1 Poiting

To examine pointing, we ran a mixed effects logistic regression to predict whether speakers chose to point during a given trial as a function of the target object's exposure rate during training, object instance in the game (first, second, or third), utility manipulation, and partner manipulation. Random effects terms for subject and object were included in the model. Consistent with our predictions, participants gestured more for words that received less training, with partners who had less knowledge, and in the condition where the utility of pointing was higher. The full results of this model are presented in Table S2.

Table S2: Propensity to point, specified as `point ~ exposureRate * partnersExposure + appearanceNum * partnersExposure + utilityCondition + (1 | subj) + (1 | realLabel)`.

term	estimate	std.error	z-value	p-value
intercept	2.07	0.35	5.96	$< .001$
exposure rate	-0.50	0.04	-12.88	$< .001$
partner's exposure	-0.81	0.22	-3.61	$< .001$
instance	0.03	0.06	0.56	.576
lower speech efficiency condition	1.20	0.32	3.79	$< .001$
partner's exposure * exposure rate	-0.21	0.03	-6.69	$< .001$
partner's exposure * instance	0.07	0.04	1.53	.127

S2.2.2 Speaking

To examine speaking, we ran a mixed-effects logistic regression to predict whether speakers chose to speak during a given trial as a function of the target object's exposure rate during training, object instance in the game (first, second, or third), utility manipulation, and partner manipulation. Random effects terms for subject and object were included in the model. Consistent with our predictions, participants used labels more for words that received more training, with partners who had more knowledge, and in the condition where the utility of speech was higher. The full results of this model are presented in Table S3.

S2.2.3 Teaching

To examine teaching, we ran a mixed-effects logistic regression to predict whether speakers chose to teach during a given trial as a function of the target object's exposure rate during training, object instance in the game (first, second, or third), utility manipulation, and partner manipulation. Random effects terms for subject and object were included in the model. Consistent with our predictions, participants taught labels more often (i.e. used both the pointing and speaking strategy simultaneously) for words that received more training, with partners who had less knowledge, and in the condition where the utility of speech was higher. The full results of this model are presented in Table S4.

Table S3: Propensity to use labeling as a strategy, specified as `speak ~ exposureRate * partnersExposure + appearanceNum * partnersExposure + utilityCondition + (1 | subj) + (1 | realLabel)`.

term	estimate	std.error	z-value	p-value
intercept	-3.19	0.30	-10.66	$< .001$
exposure rate	0.35	0.04	9.76	$< .001$
partner's exposure	1.95	0.19	10.02	$< .001$
instance	0.72	0.06	12.93	$< .001$
lower speech efficiency condition	-0.87	0.25	-3.42	.001
partner's exposure * exposure rate	0.27	0.03	9.05	$< .001$
partner's exposure * instance	-0.48	0.04	-11.07	$< .001$

Table S4: Propensity to use teaching as a strategy, specified as `teach ~ exposureRate * partnersExposure + appearanceNum * partnersExposure + utilityCondition + (1 | subj) + (1 | realLabel)`.

term	estimate	std.error	z-value	p-value
intercept	0.07	0.29	0.25	.799
exposure rate	0.14	0.04	3.21	.001
partner's exposure	-2.23	0.27	-8.38	$< .001$
instance	-1.09	0.07	-14.69	$< .001$
lower speech efficiency condition	-0.96	0.29	-3.29	.001
partner's exposure * exposure rate	-0.14	0.05	-2.70	.007
partner's exposure * instance	0.48	0.08	5.86	$< .001$

S3 Estimating model parameters

In order to explain how people choose their communicative modality in our reference game, we developed a model of communication as planning under uncertainty. This model has two parameters that need to be specified in order to get quantitative predictions out of it: (1) A rationality parameter α that controls how sensitive the model is to differences in utility, and (2) a discounting parameter γ that controls how much less the model cares about the utility it would get on its next trial than on the current trial when it makes its plan.

One way of determining the values for these parameters would be to use a Maximum Likelihood estimator to find the unique values that make the data most likely (provide the best fit). However, these point estimates would not give us a way of quantifying our uncertainty about the right parameters, nor would they correctly penalize the model correctly if only a very narrow range of the possible parameter values gave a good account for the data (see e.g. Pitt & Myung, 2002). Instead, we used Empirical Bayesian methods to estimate the full posterior probabilities of model parameters conditioned on both parameter priors and the observed data. The values reported in the paper are estimated means for these posterior distributions.

Ideally, we would like to condition on all of the data from all of the participants simultaneously. However, because of variability in the parameter values that gave good fits for different participants, the estimation algorithm did not converge in a tractable number of samples when all of the data points were considered simultaneously. To address this, we instead chose to estimate the model parameters hierarchically. We defined the generative process as one in which participants have their own individual α and γ parameters drawn from a single population α and γ distribution respectively as in the equation below (like random effects in a mixed-effects model). These group-level level rationality and discounting parameters are reported in the main text.

$$
\begin{aligned}
P(\alpha, \gamma | D) &\propto P(D|\alpha, \gamma) P(\alpha) P(\gamma) \\
&\propto \left(\prod_i P(D_i|\alpha_i, \gamma_i) P(\alpha_i|\alpha) P(\gamma_i|\gamma) \right) P(\alpha) P(\gamma)
\end{aligned}
$$

Unfortunately, estimating this full hierarchical model also proved intractable. Consequently, we first estimated each participant's individual parameters from their data using 200 steps of Metropolis-Hastings sampling after 100 burn-in samples. We used a Uniform $(0, 20)$ prior for α and a Uniform $(0, 1)$ prior for γ. The goal of these Uniform priors was to minimally impact the inferred parameter values as the true priors for them should be the population level distributions. Figure S1 shows the distribution of posterior means α and γ estimated for for each of the 401 participants.

We used the posterior mean of each participant's parameter distributions, as well as the likelihood of observing each participant's data given these parameters to estimate the population-level parameters. To do this we used 10,000 steps of Metropolis-Hastings sampling after 2,000 burn-in samples. To help the sampler converge, instead of using the prior distribution as the proposal function, we used a Normal distribution centered at the last sample with a standard deviation if .1.

At this group-level, we used theoretically-motivated priors. The rationality parameter α controls the sensitivity of the model to differences in utility and can range over the non-negative real numbers $[0, \infty)$. A rationality value of 1 corresponds to probability matching–in which actions are chosen

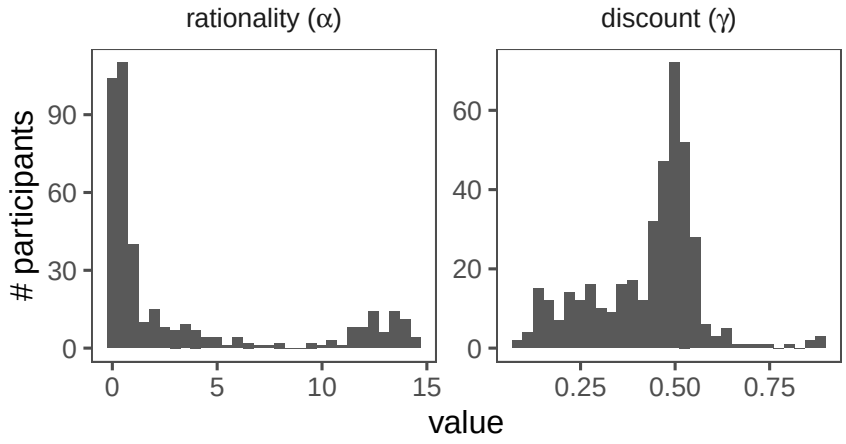

Figure S1: Estimated rationality and discount parameters for individual participants.

according to the Luce (1959) Choice rule, and larger values correspond to increasingly great moves towards maximizing. For this parameter, we chose a distribution of the prior with a mean of 1, but an Exponential shape with a long right tail. This distribution makes very high levels of rationality unlikely but not impossible:

$$P(\alpha) \sim \text{Gamma}(1, 1)$$

The discounting parameter γ controls how much the model values utility it will gain on the next step in its plan relative to the current step. This parameter can take all values between 0 and 1. As a mathematical convenience, we modeled the log odds ratio between the next trial and the previous trial with the logit function rather than assigning this parameter a prior directly. Like the squashing function in neural network models, or the linking function in logistic regression, this allows us to instead sample from all real-number values and then transform to the $(0, 1)$ range. We thus used a Normal distribution with a mean of 0 in logit space, equivalent to a γ of .5.

$$P(\mathit{logit}(\gamma)) \sim \text{Normal}(0, 1)$$

Importantly, because the model's parameters were estimated from many trials of behavior from hundreds of participants, the priors are dwarfed by the likelihoods in estimating the parameters' posterior distributions. The estimates reported in the paper are thus robust to many other prior choices. This sampling procedure was identical for all models described in the main paper. All code is available in the project page: https://osf.io/d9gkw/, as are all the samples generated in these sampling procedures.

References

Luce, R. D. (1959). *Individual choice behavior.*

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*(10), 421–425.