Integrative modeling of children's information integration during pragmatic word learning

Manuel Bohn[1], Louisa Schmidt[2], Cornelia Schulze[2], Michael C. Frank[3], & Michael Henry Tessler[4,5]

[1] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2] Leipzig Research Center for Early Child Development, Leipzig University, Leipzig, Germany

[3] Department of Psychology, Stanford University, Stanford, USA

[4] DeepMind, London, UK

[5] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, USA

Author Note

M. Bohn received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 749229. M. H. Tessler was funded by the National Science Foundation SBE Postdoctoral Research Fellowship Grant No. 1911790. M. C. Frank was supported by a Jacobs Foundation Advanced Research Fellowship and the Zhou Fund for Language and Cognition. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

The authors made the following contributions. Manuel Bohn: Conceptualization, Methodology, Formal Analysis, Visualization, Writing – original draft, Writing – review & editing; Louisa Schmidt: Conceptualization, Methodology, Investigation, Writing – review & editing; Cornelia Schulze: Conceptualization, Methodology, Writing – review & editing; Michael C. Frank: Conceptualization, Writing – review & editing; Michael Henry Tessler: Conceptualization, Methodology, Formal Analysis, Writing – review & editing.

Correspondence concerning this article should be addressed to Manuel Bohn, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. E-mail: manuel_bohn@eva.mpg.de

Abstract

150 words

Keywords: Pragmatics, language development, individual differences, cognitive modeling

Word count: X

Integrative modeling of children's information integration during pragmatic word learning

Integrative modeling

Hofman et al. (2021)

Cognitive models for theory building

Rooij (2022)

Guest and Martin (2021)

Rooij and Baggio (2021)

In developmental psychology

Simmering, Triesch, Deák, and Spencer (2010)

Ullman and Tenenbaum (2020)

In pragmatic language comprehension

Bohn, Tessler, Merrick, and Frank (2022)

Tessler and Goodman (2019)

Using RSA to study individual differences:

Franke and Degen (2016)

Bohn, Tessler, Kordt, Hausmann, and Frank (2022)

Pragmatics tasks show good re-test reliability.

Part 1: Sensitivity

Methods

Methods, sample size and analyses were pre-registered at: https://osf.io/pa5x2. All data, analysis scripts, model code and experimental procedures are publicly available in the following online repository: https://github.com/manuelbohn/spin-within.

Participants. We collected complete data for 60 children ($m_{age} = 4.11$, range$_{age}$: 3.06 - 4.93, 30 girls). In addition ... [Louisa - könntest Du das ergänzen]. Children came from an ethnically homogeneous, mid-size German city (~550,000 inhabitants, median income €1,974 per month as of 2020); were mostly monolingual and had mixed socioeconomic backgrounds. The study was approved by an internal ethics committee at the Max Planck Institute for Evolutionary Anthropology. Data was collected between ... [Louisa].

Procedure. Children were recruited via a database and participated with their parents via an online conferencing tool. The different tasks were programmed as interactive picture books in `JavaScript/HTML` and presented on a website. During the video call, participants would enter the website with the different tasks and share their screen. The experimenter guided them through the procedure and told caregivers when to advance to the next task. Children responded by pointing to objects on the screen, which their caregivers would then select for them via mouse click. For the production task, the experimenter shared their screen and presented pictures in a slide show. For the mutual exclusivity, discourse novelty, and combination tasks, pre-recorded sound files were used to address the child. Figure 1 shows screenshots from the different tasks.

In the *discourse novelty* task, children saw a speaker (cartoon animal) standing between two tables. On one table, there was a novel object (drawn for the purpose of this study) while the other was empty. The speaker sequentially turned to both sides (order counterbalanced) and either commented on the presence or absence of an object (without using any labels). Then, the speaker disappeared and – while the speaker was gone – another novel object appeared on the previously empty table. Next, the speaker re-appeared and requested one of the objects using a novel non-word as the label. We assumed that children would take the novel word to refer to the object that was new to the speaker. Children received 16 trials, each with a new pair of novel objects. The location of the empty table was counterbalanced.

In the *mutual exclusivity* task, children again saw a speaker and two tables. On one table, there was a novel object while on the other there was a (potentially) familiar object. The speaker used a novel non-word to request one of the objects. We assumed that children would take the novel word to refer to the novel object. In line with previous work (Bohn, Tessler, Merrick, & Frank, 2021; Grassmann, Schulze, & Tomasello, 2015; Lewis, Cristiano, Lake, Kwan, & Frank, 2020) we assumed this inference would be modulated by children's lexical knowledge of the familiar object. Children received 16 trials, each with a new pair of novel and familiar objects. The location of the familiar object was counterbalanced.

In the *word production* task, the experimenter showed the child each of the 16 familiar objects from the mutual exclusivity task and asked them to name it. We used a pre-defined list of acceptable labels per object to categorize children's responses as either correct or incorrect.

In the *word comprehension* task, the child saw four slides with six objects. Four objects per slide were taken from the 16 familiar objects that also featured in the mutual exclusivity and word production tasks. Two objects were unrelated distractors. The experimenter labelled one familiar object after the other and asked the child to point to it.

Data collection was split into two sessions scheduled for two consecutive?? [Louisa] days. On day one, children completed the mutual exclusivity and the discourse novelty tasks. On day two, they completed the combination task followed by the word comprehension and production tasks.

Analysis

The focus of the analysis was on estimating person-specific parameters for each inforamtion source. Models to estimate parameters were implemented in the probabilistic programming language `webppl` (Goodman & Stuhlmüller, 2014). The three information sources were: sensitivity to common ground (ρ_i), expectations about speaker informativeness (α_i), and semantic knowledge (θ_{ij}). Figure 1 shows which tasks informed which parameters.

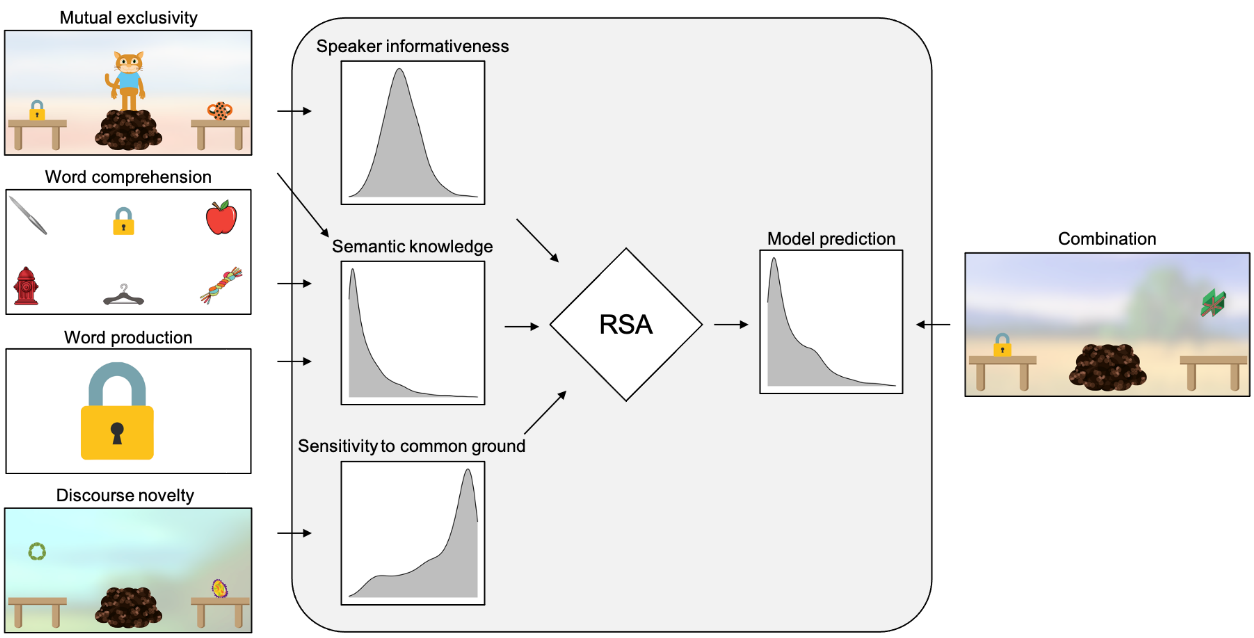

Figure 1. Schematic overview of the study and the model. Pictures on the left show screenshots from the four sensitivity tasks. Arrows indicate which tasks informed which parameter in the model (grey area). Based on the data from the sensitivity tasks, child specific parameter distributions for each information source were estimated. These sources were integrated via an RSA model, which generated predictions for each trial of the combination task. These predictions were then evaluated against new data from the combination task.

All parameters were estimated via hierarchical regression (mixed-effects) models. That is, for each parameter, we estimated an intercept and slope (fixed effects) that best described the developmental trajectory for this parameter based on the available data. Participant-specific parameters values (random effects) were estimated as deviations from the value expected for a participant based on their age. Details about the estimation procedure can be found in the supplementary material. The code to run the models can be found in the associated online repository.

The parameters for semantic knowledge (θ_{ij}) were simultaneously informed by the data from the mutual exclusivity, the comprehension and the production experiments. To leverage the mutual exclusivity data, we adopted the RSA model described in Part 2 to a situation in which both objects (novel and familiar) had equal prior probability (i.e., no common ground information). In the same model, we also estimated the parameter for speaker informativeness (see below). For the comprehension experiment, we simply assumed that the child was able to select the correct word with probability θ_{ij}. If the child did not know the word, we assumed they would select the correct word at a rate expected by chance (1/6). For the production experiment, we assumed that if the child knew the word (a function of θ_{ij}), they produced the word with probability γ. This successful-production-probability γ was the same for all children and was inferred based on the data. This adjustment reflects the finding that children's receptive vocabulary for nouns tends to be larger than the productive (Clark & Hecht, 1983; Frank, Braginsky, Yurovsky, & Marchman, 2021). Taken together, for each child i and familiar object j there were three data points to inform θ: one trial from the mutual exclusivity, one from the comprehension and one from the production experiment.

The parameter representing a child's expectations about how informative speakers are (α_i), was estimated based on the data from the mutual exclusivity experiment. As mentioned above, this was done jointly with semantic knowledge in a RSA model adopted to a situation with equal prior probability of the two objects (novel and familiar). Thus, for each child, there were 16 data points to inform α.

We estimated children's sensitivity to common ground (ρ_i) based on the data from the discourse novelty experiment. This was done via simple logistic regression and based on the 12 data points from this task.

Results

Figure 2 visualizes the results for the four sensitivity tasks and the person specific model parameters estimated from the data. In all four tasks, we saw that children performed above chance (not applicable in the case of word production), suggesting that they made the alleged pragmatic inference or knew (some) of the words for the objects involved. With respect to age, performance in raw test scores seemed to increase with age in the three tasks relying on semantic knowledge (mutual exclusivity, word production and word comprehension). Performance in these tasks was also correlated (see supplementary material). For discourse novelty, performance did not increase with age. Most importantly, however, we saw considerable variation between individuals. When focusing on the individual-specific parameter estimates (Figure 2B), we saw that parameters that were estimated based on more data (sensitivity to common ground – 12 trials, and expectations about speaker informativeness – 16 trials) had better defined posterior distributions compared to semantic knowledge (3 trials per object).

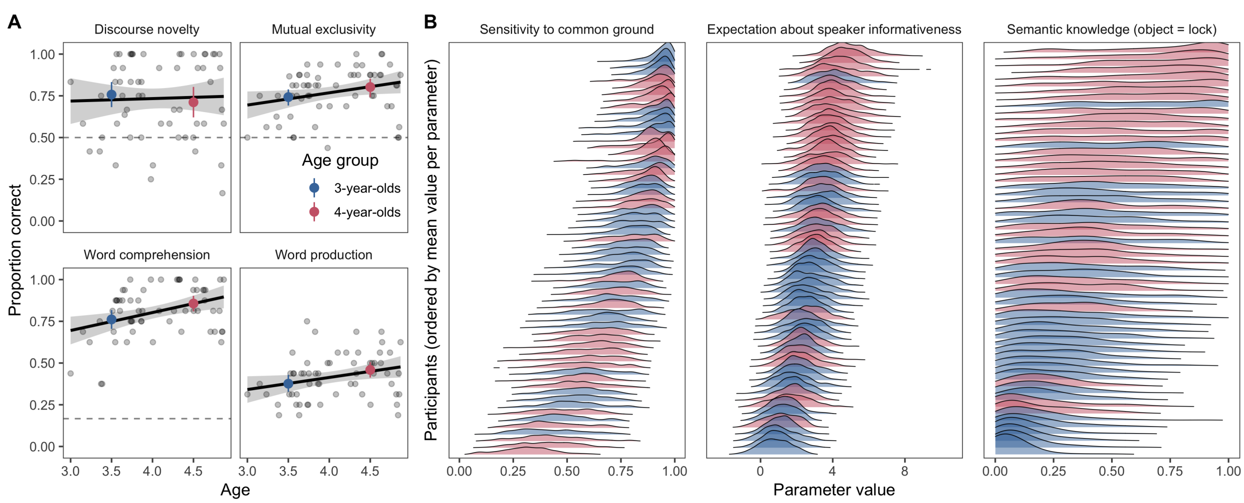

Figure 2. Results for the sensitivity tasks. A: proportion of correct responses in each task by age. Colored dots show the mean proportion of correct responses (with 95% CI) binned by year. Regression lines show fitted generalized linear models with 95% CIs. B: posterior distributions for each parameter (information source) and participant, ordered by mean value, separate for each parameter. Color shows age group.

Discussion

The goal of Part 1 was to estimate person-specific parameters representing each individual's sensitivity to the three information sources. We found that, as a group, children were sensitive to the different information sources. Furthermore, there was substantial variation between individuals in *how* sensitive they were to each information source. These results provided a solid basis for studying information integration in Part 2.

Part 2: Integration

In Part 2, we studied how children integrate the three information sources. We incorporated the parameters estimated in Part 1 in a computational cognitive model of pragmatic reasoning to generate participant-specific predictions about how the three information sources should be integrated. We then compared these predictions to new data collected with a task in which all three information sources were manipulated. We used Bayesian model comparisons to compare our focal *rational integration model* to alternative models that made different theoretical assumptions about the integration process.

Methods

The study was pre-registered and all data, analysis script and materials are publicly available (see Part 1 for more information).

Participants. Participants were the same as in Part 1.

Procedure. The task was implemented in the same environment as the tasks in Part 1. Each child completed the combination task on the second testing day. The general procedure followed that of the novelty task, however, only one of the objects was unknown while the other was familiar. The combination task had two conditions. In the *congruent condition*, the object that was new to discourse was the novel object. As a consequence, mutual exclusivity and discourse inferences pointed to the same object as the referent of the novel word were aligned. In the *incongruent condition*, the familiar object was new to discourse and thus, the two inferences pointed to different objects. We created matched pairs for the 16 familiar objects and assigned one object of each pair to one of the two conditions.

Thus, there were eight trials per condition in the combination task in which each trial was with a different familiar object. We counterbalanced the order of conditions and the side on which the discourse-novel object appeared. Responses were coded from a mutual exclusivity perspective (choosing novel object = 1). All children received the same order of trials. There was the option to terminate the study after 8 trials (two children).

Analysis

We adopted the modelling framework used by Bohn et al. (2021). Our models are situated in the Rational Speech Act (RSA) framework (Frank & Goodman, 2012; Goodman & Frank, 2016). RSA models treat language understanding as a special case of Bayesian social reasoning. A listener interprets an utterance by assuming it was produced by a cooperative speaker who has the goal to be informative. Being informative is defined as producing messages that increase the probability of the listener inferring the speaker's intended message. The focal *rational integration* model, including all data-analytic parameters, is formally defined as:

$$P_{L_1}(r \mid u; \{\rho_i, \alpha_i\, \theta_{ij}\}) \propto P_{S_1}(u \mid r; \{\alpha_i, \theta_{ij}\}) \cdot P(r \mid \rho_i) \tag{1}$$

The model describes a listener (L_1) reasoning about the intended referent of a speaker's (S_1) utterance. This reasoning is contextualized by the prior probability of each referent $P(r \mid \rho_i)$. This prior probability is a function of the common ground ρ shared between speaker and listener in that interacting around the objects changes the probability that they will be referred to later.

To decide between referents, the listener (L_1) reasons about what a rational speaker (S_1) would say given an intended referent. This speaker is assumed to compute the informativity for each available utterance and then choose the most informative one. The expectation of speaker informativeness may vary and is captured by the parameter α:

$$P_{S_1}(u \mid r; \{\alpha_i \theta_{ij}\}) \propto P_{L_0}(r \mid u; \{\theta_{ij}\})^{\alpha_i} \tag{2}$$

The informativity of each utterance is given by imagining which referent a literal listener (L_0), who interprets words according to their lexicon $\mathcal{L}$, would infer upon hearing the utterance. This reasoning depends on what kind of semantic knowledge (word–object mappings, $\theta_{i,j}$) the speaker thinks the literal listener has. As noted above, for familiar objects, we take semantic knowledge to be a function of the degree-of-acquisition of the associated word.

$$P_{L_0}(r \mid u; \{\theta_{ij}\}) \propto \mathcal{L}(u, r \mid \theta_{ij}) \tag{3}$$

This modelling framework allows us to generate predictions for each participant and trial in the combination task based on the participant-specific parameters estimated in Part 1. That is, for each combination of ρ, α, and θ for participant i and familiar object j, the model returns a distribution for the probability with which the child should choose the novel object. We contrasted the predictions made by the *rational integration* model described above to those made by two plausible alternative models which assume that children selectively ignore some of the available information sources (Gagliardi, Feldman, & Lidz, 2017). These models generated predictions based on the same parameters as the *rational integration* model, the only difference lay in how the parameters were used. The *no speaker informativeness* model assumed that the speaker does not communicate in an informative way and therefore focused on the sensitivity to common ground. The *no common ground* model ignores common ground information and focused on the mutual exclusivity inference (speaker informativeness and semantic knowledge instead). A detailed description of all the models along with technical information about parameter estimation can be found in the supplementary material.

We evaluated the model predictions in two steps. First, we replicated the group-level results of Bohn et al. (2021). That is, we compared the three models in how well they predict the data of the combination task when aggregating across individuals. For this, we correlated model predictions and the data (aggregated by trial and age group) and computed pairwise Bayes Factors based on the marginal likelihood of the data given the model.

Second, and most importantly, we evaluated how well the model predicted performance on an *individual* level. For each trial, we converted the (continuous) probability distribution returned by the model into a binary prediction (the structure of the data) by flipping a coin with the Maximum a posteriori estimate (MAP) of the distribution as its weight. For the focal and the two alternative models, we then computed the proportion of trials for which the model predictions matched children's responses and compared them to a level expected by random guessing using a Bayesian t-test. Finally, for each child, we computed the Bayes Factor in in favor of the *rational integration* model and checked for how many children this value was above 1 (log-Bayes Factors > 0). Bayes Factors larger than 1 present evidence in favor of the *rational integration* model. We evaluated the distribution of Bayes Factors following the classification of Lee and Wagenmakers (2014).

Results

On a group-level, the results of the present study replicated those of Bohn et al. (2021). The predictions made by the *rational integration* model were highly correlated with children's responses in the combination task. The model explained around 74% of the variance in the data and with that more compared to the two alternative models (Figure 3A). Bayes Factors computed via the marginal likelihood of the data (Figure 3B) strongly favored the *rational integration* model in comparison to the *no common ground* (BF_{10} = 9.1e+53) as well as the *no speaker informativeness* model (BF_{10} = 1.2e+44).

Finally, we turned to the individual-level results. When looking at the proportion of correct predictions, we saw that the *rational integration* model correctly predicted children's

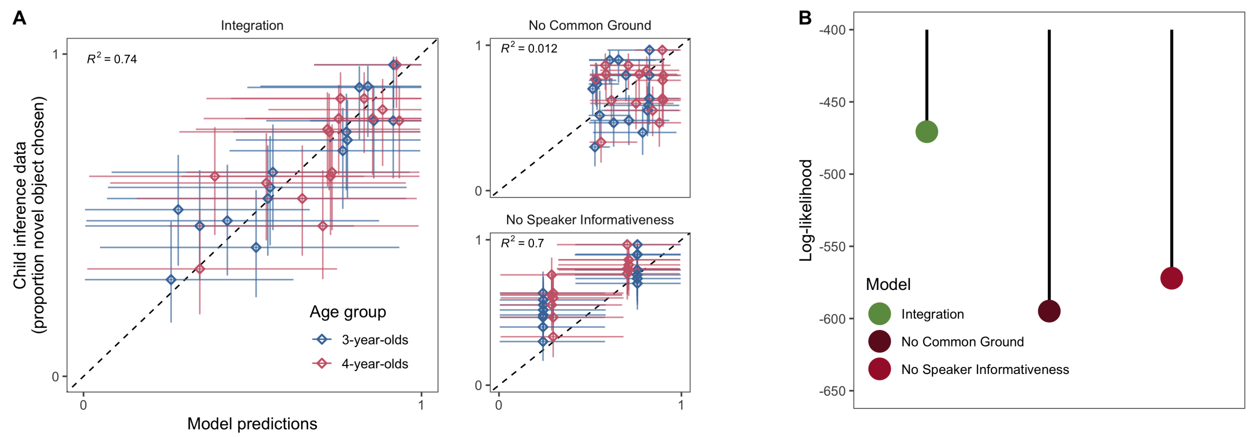

Figure 3. Group-level model comparison. A: Correlation between model predictions and data (aggregated across individuals and binned by year with 95%HDI) for each trial in the combination experiment. B: log-likelihood for each model given the data.

responses in the combination task in 72% of trials, which was well above chance (BF_{10} = 2.15e+14) and higher compared to the two alternative models (Figure 4A). Note that the alternative models also predicted children's responses at a level above chance (*no common ground*: 61%, $BF_{10} = 220251$; *no speaker informativeness*: 60%, $BF_{10} = 55.4$), emphasizing that they constitute plausible alternatives. In the supplementary material we also compared models with respect to the situations in which they did or did not correctly predict children's responses.

When directly comparing the models on an individual level, we found that the *rational integration* model provided the best fit for the majority of children. In comparison to the *no common ground* model, 62% of Bayes Factors were larger than 1 and 35% were larger than 10. In comparison to the *no speaker informativeness* model, 68% of Bayes Factors were larger than 1 and 45% were larger than 10 (Figure 4B).

Discussion

The results of Part 2 show that the *rational integration* model accurately predicted children's responses in the combination task. Importantly, this was the case not just on a

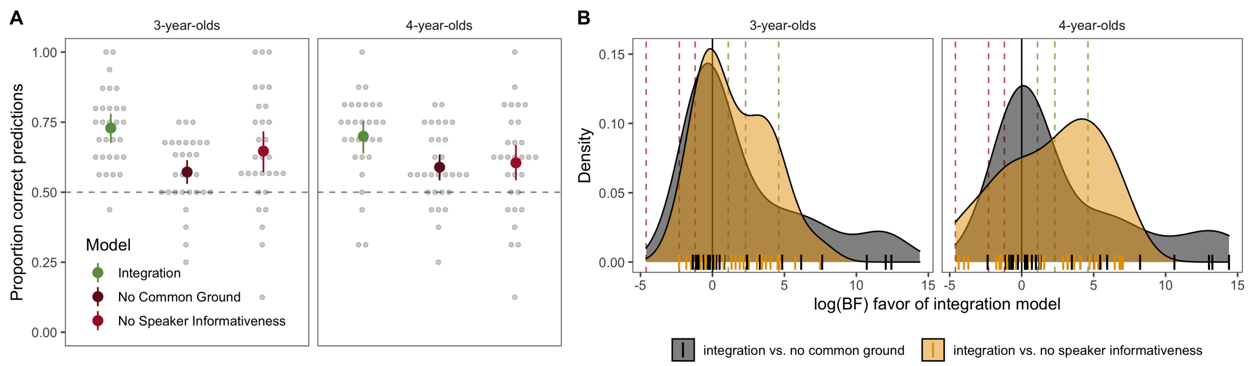

Figure 4. Individual-level model comparison. A: proportion of correct predictions for each model. Colored dots show mean with 95%CI. Light dots show aggregated individual data. B: distribution of log-Bayes Factors for each individual. Dashed lines show Bayes Factor thresholds of 3, 10 and 100.

group level, but also on an individual level. Based on the sensitivity measures obtained for each child in Part 2, the model correctly predicted children's responses in the majority of trials. Furthermore, it was more likely to be correct and provided a better explanation of the data compared to two alternative models that assumed that children selectively ignored some of the information sources.

General discussion

Models work on individual level. this work shows they make good predictions and also model comparison is a great tool to contrast theories. Psychological reality of the model and their parameters are still in question, but they work well. Recent other work suggests model parameters can be used in individual differences studies, representing differences between individuals as an alternative to raw scores. Allows linking different paradigms on a process level.

References

Bohn, M., Tessler, M. H., Kordt, C., Hausmann, T., & Frank, M. C. (2022). *An individual differences perspective on the development of pragmatic abilities in the preschool years*.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour*, *5*(8), 1046–1054.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2022). Predicting pragmatic cue integration in adults' and children's inferences about novel word meanings. *Journal of Experimental Psychology: General*.

Clark, E. V., & Hecht, B. F. (1983). Comprehension, production, and language acquisition. *Annual Review of Psychology*, *34*(1), 325–349.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

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

Franke, M., & Degen, J. (2016). Reasoning in reference games: Individual-vs. Population-level probabilistic modeling. *PloS One*, *11*(5), e0154854.

Gagliardi, A., Feldman, N. H., & Lidz, J. (2017). Modeling statistical insensitivity: Sources of suboptimal behavior. *Cognitive Science*, *41*(1), 188–217.

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. (2014). *The design and implementation of probabilistic programming languages*. http://dippl.org.

Grassmann, S., Schulze, C., & Tomasello, M. (2015). Children's level of word knowledge predicts their exclusion of familiar objects as referents of novel words. *Frontiers in Psychology*, *6*, 1200.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ... others. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, *198*, 104191.

Rooij, I. van. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, 1–2.

Rooij, I. van, & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, *16*(4), 682–697.

Simmering, V. R., Triesch, J., Deák, G. O., & Spencer, J. P. (2010). A dialogue on the role of computational modeling in developmental science. *Child Development Perspectives*, *4*(2), 152–158.

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological Review*, *126*(3), 395.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533–558.