What's she talking about? Category based discourse inferences in early childhood

Manuel Bohn[1,2], Khuyen Nha Le[1], Benjamin Peloquin[1], Bahar Köymen[3], & Michael C. Frank[1]

[1] Stanford University

[2] Leipzig University

[3] University of Manchester

Author Note

We thank Megan Merrick and Sabina Zacco for their help with the data collection. MB received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 749229.

Correspondence concerning this article should be addressed to Manuel Bohn, Jahnallee 59, 04109 Leipzig, Germany. E-mail: manuel.bohn@uni-leipzig.de

Abstract

tbd...

Keywords: keywords

Word count: X

What's she talking about? Category based discourse inferences in early childhood

Experiment 1

All experimental procedures, sample sizes and statistical analysis were pre-registered (see https://osf.io/9ypxn and https://osf.io/fyaxq). The experimental procedure can be found in the assoicated online repository at https://github.com/manuelbohn/disCon.

Participants

We obtained valid data from 71 children, including 30 2-year-olds (mean = 2.63, range = 2.00 - 2.98), 21 3-year-olds (mean = 3.56, range = 3.13 - 3.97) and 20 4-year-olds (mean = 4.50, range = 4.00 - 4.97). We tested a larger sample of 2-year-olds because we expected a weaker effect in this age group. In addition, 12 children were recruited but not tested becasue their parents reported less than 75% of English exposure at home. Ten children started the experiment but did not finish it because they became impatient (7) or the equipment broke (3). Three children were tested but excluded becasue they were correct in less than 5/6 training trials (see below). All children were recruited from the floor of a Children's museum in San José, California, USA. The population from which this sample is drawn is characterised by diverse ethnic background and high socioeconimic status. Parents gave informed consent and provided demographic information. All experiments reported in this paper were approved by the Stanford Institutional Review Board (protocol no. 357 19960).

Method

Study materials were presented as a picture book on a tablet computer (Frank, Sugarman, Horowitz, Lewis, & Yurovsky, 2016). Children reponded by touching objects on the screen. Responses were automatically saved. The experimenter guided children through

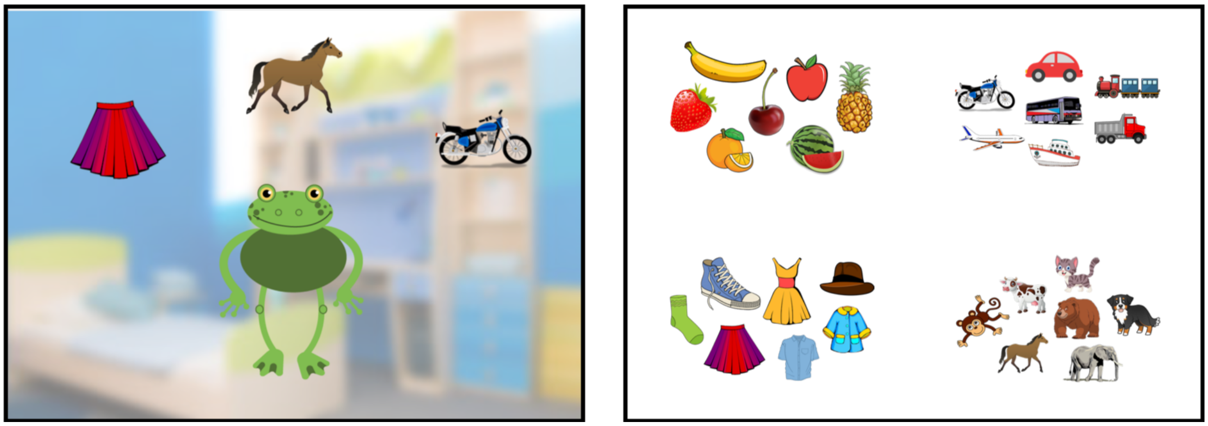

Figure 1. Left: Screenshot from the experimental setup. Right: Stimuli pictures for the four categories: fruits, vehicles, clothes and mammals.

the procedure and read out general instructions. The study was framed as visit to the house of the little animals which would show the child the things they have at home. Utterances made by the different animals were pre-recorded from native English speakers, with one speaker per animal. On each trial, children saw one animal in the middle of the screen with three objects above them (Figure 1, left). Each objects was from different category (mammals, vehicles, clothes and fruits). For each category, we had pictures of seven different category memebers (e.g. for vehicles: car, truck, train, bus, airplane, boat and motorbike, see Figure 1, right). The trial started with six training rounds, in which the animal named one of the objects above them, asking the child to touch it (e.g. "Look at that, can you touch the horse"). From one round to the next, the pictures changed but the categories remained the same. For example, children saw a skirt, a horse and a motorcylce on the first training round and a jacket, a dog and a bus on the second. During training, the speaker consitently named objects from the same target category. After six training rounds, children received a test round in which the speaker used a pronoun to refer to one of the objects ("Look at that, can you touch *it*"). Categories were ranomdly selected at the beginning of each trial and so was the order of pictures within each category. The position of each picture (left, right mddle) was also randomly determined on each round. Children received four trials, one with each category as the target.

Results

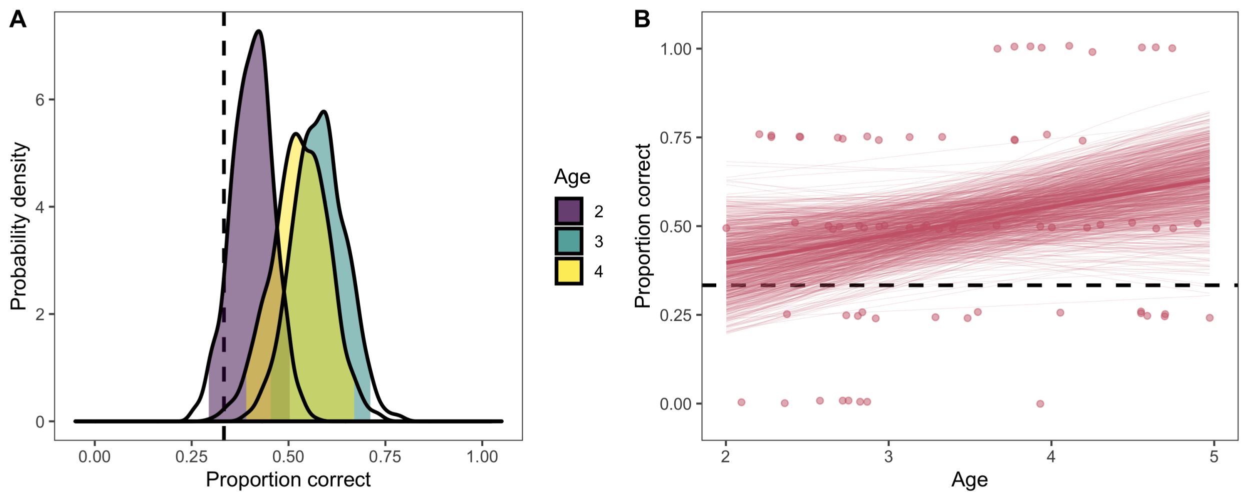

Figure 2. A: Posterior probability distribution for the mean for each age bin. Shaded regions indicate 95% CIs. B: Correct responses for age continously. Transparent dots show aggregated data from individual participants. Red line with grey shows the smoothed conditional mean of the data in each condition..

The dependent variable in all analysis was whether the touched object at test was from the same category as the objects named throughout the training rounds. All following analysis were computed in R (R Core Team, 2018). As a first step, we aggregated respones across trials for each child and compared the proportion of correct responses to a level expected by chance (33% correct) within each age bin. We used the function `ttestBF` from the R-package `BayesFactor` (Morey & Rouder, 2018) to compute a Bayes factor (BF) in favor of the hypothesis that performance is above chance (see Figure 2A). We found little evidence that 2-year olds performed above chance (mean proportion correct = 0.42, BF = 0.59) but found subtantial evidence for 3-year-olds (mean proportion correct = 0.60, BF = 90.77) and 4-year-olds (mean proportion correct = 0.55, BF = 10.39).

To analyse responses continously across age we used generalized linear mixed models (GLMM) fit via the function `brm` from the R-package `brms` (Bürkner, 2017). All models had default priors and included random effects for participant id and speaker. Inference was based on comparing models that differed in whether they included the key predictor of interest, in this case age. Following McElreath (2016), we compared models using WAIC (widely applicable information criterion) scores and weights. WAIC is an indicator of the model's predictive accuracy for out of sample data and model's with lower scores are preferred. WAIC weights are an estimate of the probability that this model (compared to all other models considered) will make the best predictions on new data. We quantified the evidence in favor of the model with the lowest WAIC score (highest WAIC weight) compared to alternative models by computing Bayes factors via the funtion `bayes_factor` from the R-package `brms` (Bürkner, 2017).

The model comparison favored the model including age as a predictor (Table 1). The model estiamte for age was positive ($\beta = 0.32$, 95% confidence interval (CI) = -0.08 - 0.75), suggesting an increase in performance with age (see also Figure 2B). However, the evidence in support of this model was modest, speaking against subtantial developmental gains across the age range considered.

Discussion

Experiment 2

Registration: https://osf.io/x2k4p

Table 1

Model comparison for Experiment 1

Model	WAIC	SE	weight	BF
correct ~ age + RE	387.20	7.67	0.56	-
correct ~ 1 + RE	387.69	7.11	0.44	1.77

Note. All models had the same random effects (RE) structure. BF denotes the Bayes Factor in favor to the model with the highest WAIC weight.

Participants

Material and Procedure

Results and Discussion

Experiment 3

Registration: https://osf.io/5e9pk

Participants

Material and Procedure

Results and Discussion

General Discussion

Table 2

Model comparison for Experiment 2

Model	WAIC	SE	weight	BF
correct ~ age + RE	176.48	9.28	0.43	-
correct ~ age * condition + RE	177.29	11.00	0.29	0.09
correct ~ age + condition + RE	177.36	9.93	0.28	0.45

Note. All models had the same random effects (RE) structure. BF denotes the Bayes Factor in favor to the model with the highest WAIC weight.

Table 3

Model comparison for Experiment 3

Model	WAIC	SE	weight	BF
correct ~ age * condition + RE	325.21	10.50	0.58	-
correct ~ age + condition + RE	327.25	9.52	0.21	23.91
correct ~ age + RE	327.30	8.96	0.21	23.88

Note. All models had the same random effects (RE) structure. BF denotes the Bayes Factor in favor to the model with the highest WAIC weight.

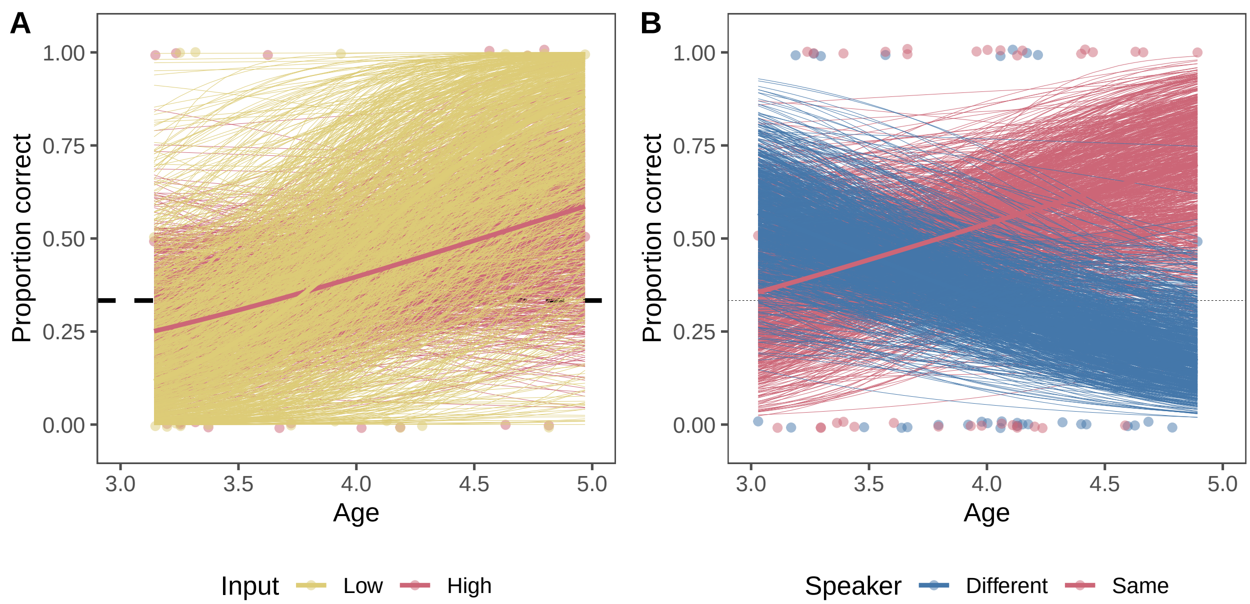

Figure 3

References

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi:10.18637/jss.v080.i01

Frank, M. C., Sugarman, E., Horowitz, A. C., Lewis, M. L., & Yurovsky, D. (2016). Using tablets to collect data from young children. *Journal of Cognition and Development*, *17*(1), 1–17.

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan* (pp. xvii, 469). Boca Raton: CRC Press.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

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