Category based discourse inferences in early childhood Supplementary material

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

Experiment 1

Method

Materials. The list below shows all words that were used to request objects during training rounds for each of the four categories. The corresponding pictures are shown in figure 1 in the main manuscript.

- fruit: strawberry, apple, banana, cherry, orange, melon, pineapple
- vehicles: car, truck, train, bus, airplane, boat, motorbike
- clothes: shoe, sock, hat, shirt, jacket, dress, skirt
- animals: dog, cat, horse, bear, cow, monkey, elephant

The object shown at test for each category was randomly selected from all objects of that category.

Table S1
Bayes factors based on Bayesian t-test with different priors on standardized effect size

Age	default	wide	ultrawide
2	0.59	0.45	0.34
3	90.77	90.58	81.99
4	10.39	9.52	8.03

Analysis. We used R (Version 3.6.1; R Core Team, 2019) and the R-packages *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *brms* (Version 2.10.0; Bürkner, 2017, 2018), *broom* (Version 0.5.2; Robinson & Hayes, 2019), *coda* (Version 0.19.3; Plummer, Best, Cowles, & Vines, 2006), *dplyr* (Version 0.8.3; Wickham, François, Henry, & Müller, 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), *langcog* (Version 0.1.9001; Braginsky, Yurovsky, & Frank, 2019), *magrittr* (Version 1.5; Bache & Wickham, 2014), *Matrix* (Version 1.2.17; Bates & Maechler, 2019), *papaja* (Version 0.1.0.9842; Aust & Barth, 2018), *purrr* (Version 0.3.2; Henry & Wickham, 2019), *Rcpp* (Eddelbuettel & Balamuta, 2017; Version 1.0.2; Eddelbuettel & François, 2011), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *reshape2* (Version 1.4.3; Wickham, 2007), *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.

Results. We conducted a sensitivity analysis for the Bayes factors that informed the comparison to chance in each age group. We re-ran the analysis using wider priors than the default priors from the package. Table S1 reports the results. Results show that the conclusions - 3- and 4-year-olds, but not 2-year-olds, select the object from the same category above chance - are robust to changes in the prior width.

Supplementary Experiment

In this study, we varied the number of training rounds before the test event. Procedures and analysis were pre-registered at https://osf.io/x2k4p. Our main focus was on condition differences and we therefore sampled a smaller number of children from each age group.

Method

Participants. We tested 33 children, including 19 3-year-olds (mean = 3.45, range = 3.00 - 3.93, 5 girls) and 14 4-year-olds (mean = 4.44, range = 4.02 - 4.97, 7 girls). For details on population characteristics and ethical approval see experiment 1.

Materials and Procedure. Study material and procedure were identical to study 1 with the following change: we administered two types of trials that varied in the number of training rounds that preceded the test. *Low input* trials had one training round while *high input* trials had six training rounds. Participants received four trials, two in each condition. The order of conditions was randomized. Categories were randomly assigned to each condition and object positions were randomized in the same way as in study 1. Experimental procedures can be found in the associated online repository.

Results. We used WAIC scores and weights to compare models including condition as a predictor to models lacking it. Table S2 shows the results of the model comparison. The model without condition as a predictor (either as main effect or as an interaction with age) provided the best fit. In this model, the predictor for age was largely positive and different from zero (β = 1.06, 95% CI = 0.00 - 2.31) suggesting an increase in performance with age. Figure S1 shows that, in contrast to study 1, 3-year-olds had difficulties with this version of the task.

Discussion. The results of this experiment suggest that hearing fewer examples of objects from the implied category does not automatically result in worse performance overall. However, this interpretation should be taken with caution. As it turns out, mixing

Table S2
Model comparison for Experiment 2

Model	WAIC	SE	weight
correct ~ age + RE	176.37	9.28	0.45
correct ~ age + condition + RE	177.26	9.90	0.29
correct ~ age * condition + RE	177.40	10.93	0.27

Note. All models included random intercepts for particiapnt and speaker and random slopes for condition.

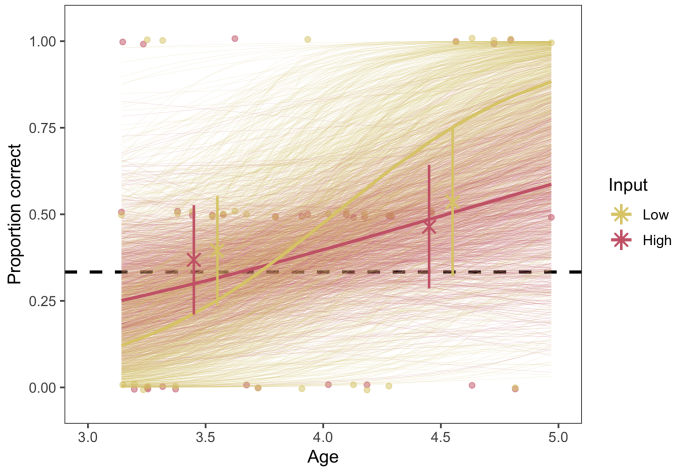

Figure S1. Correct responses in supplementary experiment by age and condition. Transparent dots show aggregated data from individual participants. Blue and red crosses show mean within age bin based on aggregated data with 95% CI based on non-parametric bootstrap. Colored lines show the mean of the posterior distribution for each condition based on the interaction model (note that the model comparison favored the model without condition). Lighter lines show 1000 random draws per condition from the model posterior to depict uncertainty. Dotted line indicates level of performance expected by chance.

low and high input trials substantially affected 3-year-olds' performance in the task. That is, in contrast to study 1 and 2, they struggled with the basic inference. This suggests that including low input trials made the task harder overall. The difference between conditions might be more prominent in a between subjects design. However, more data is needed to reach a conclusion and, as of now, we see these results as inconclusive.

Exploratory analysis

This is an exploratory analysis that was not pre-registered. From all experiments we select the data with six training rounds and the same speaker. That is, we included all data from study 1, the high input condition from the supplementary experiment and the same speaker condition from study 2.

Methods

Participants. Data from all children (N = 164) who participated in one of the three experiments were included. For detailed information about age and population characteristics see individual experiments.

Results. Figure S2A shows that performance differed widely across categories. This visual impression was corroborated in a model comparison. We compared a model with condition to a model without condition as a predictor in the same way as in previous experiments. Table S3 shows that the model including category was clearly favored. For comparisons between categories we set clothes as the reference category. Figure S2B and S3 show that children chose the object from the same category most often when fruits was the target category. Vehicles and animals (reference category) resulted in intermediate levels of performance while clothes was the most difficult one.

Discussion

This exploratory analysis suggests that children's ability to infer the topic of a conversation depends, in part, on the implied topic itself. Categories may differ in difficulty either because some category members might be less familiar to children and/or because children are less familiar with the category itself.

Table S3
Model comparison for exploratory analysis on category differences

Model	WAIC	SE	weight
correct ~ age + category + RE	628.32	13.65	0.92
correct ~ age + RE	633.29	11.44	0.08

Note. All models included random intercepts for particiapnt and speaker and random slopes for category within speaker

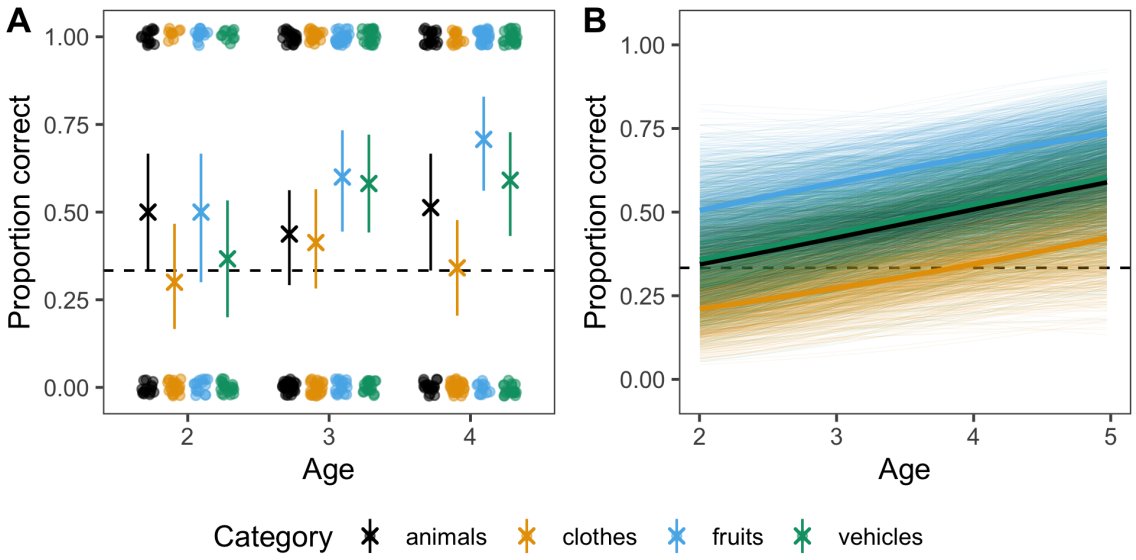

Figure S2. A: Data plotted by age bin and category. Transparent dots show data from individual participants. Colored crosses show mean within age bin and category with 95% confidence intervals based on non-parametric bootstrap. B: Correct responses for age continously. Colored lines show the mean of the posterior distribution of the model including category. Lighter lines show 1000 random draws from the model posterior to depict uncertainty in the model. Dotted line indicates level of performance expected by chance.

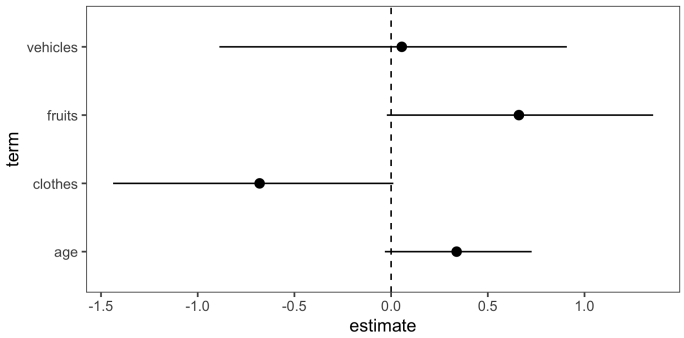

Figure S3. Model estimates for each category with clothes as the reference category. Dots show mean and error bars 95% credible intervals of the posterior distribution.

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. (2019). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Braginsky, M., Yurovsky, D., & Frank, M. (2019). *Langcog: Language and cognition lab things*. Retrieved from http://github.com/langcog/langcog

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

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. doi:10.32614/RJ-2018-017

Eddelbuettel, D., & Balamuta, J. J. (2017). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, *5*, e3188v1. doi:10.7287/peerj.preprints.3188v1

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. doi:10.18637/jss.v040.i08

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

Kassambara, A. (2019). *Ggpubr: 'Ggplot2' based publication ready plots*. Retrieved from

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

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

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

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. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Robinson, D., & Hayes, A. (2019). *Broom: Convert statistical analysis objects into tidy tibbles*. Retrieved from https://CRAN.R-project.org/package=broom

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. Retrieved from http://www.jstatsoft.org/v21/i12/

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.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., & Henry, L. (2019). *Tidyr: Tidy messy data*. 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