Supplementary material: Children's interpretation of ambiguous pronouns based on prior discourse

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

[1] Department of Psychology, Stanford University

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

[3] School of Health Sciences, University of Manchester

Experiment 1

Method

Materials. The list below shows all words that were used to request objects during training trials 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 prior	wide prior	ultrawide prior
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 3.6.3; R Core Team, 2019) and the R-packages *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *bayesplot* (Version 1.7.2; Gabry, Simpson, Vehtari, Betancourt, & Gelman, 2019), *brms* (Version 2.13.5; Bürkner, 2017, 2018), *broom* (Version 0.7.0; Robinson & Hayes, 2019), *coda* (Version 0.19.3; Plummer, Best, Cowles, & Vines, 2006), *dplyr* (Version 1.0.1; Wickham et al., 2019), *forcats* (Version 0.5.0; Wickham, 2019a), *ggplot2* (Version 3.3.2; Wickham, 2016), *ggpubr* (Version 0.4.0; Kassambara, 2019), *ggridges* (Version 0.5.2; Wilke, 2020), *ggthemes* (Version 4.2.0; Arnold, 2019), *knitr* (Version 1.29; Xie, 2015), *langcog* (Version 0.1.9001; Braginsky, Yurovsky, & Frank, 2019), *magrittr* (Version 1.5; Bache & Wickham, 2014), *Matrix* (Version 1.2.18; Bates & Maechler, 2019), *papaja* (Version 0.1.0.9997; Aust & Barth, 2018), *purrr* (Version 0.3.4; Henry & Wickham, 2019), *Rcpp* (Version 1.0.5; Eddelbuettel & François, 2011; Eddelbuettel & Balamuta, 2017), *readr* (Version 1.3.1; Wickham, Hester, & Francois, 2018), *reshape2* (Version 1.4.4; Wickham, 2007), *stringr* (Version 1.4.0; Wickham, 2019b), *tibble* (Version 3.0.3; Müller & Wickham, 2019), *tidyr* (Version 1.1.1; Wickham & Henry, 2019), and *tidyverse* (Version 1.3.0; 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.

Experiment 1 - adults

Method

We tested 20 adults recruited via Amazon Mechanical Turk (MTurk). Participants were restricted to have an US IP address and received payment equivalent to an hourly wage of ~ $9.

Materials

The setup and the procedure was the same as for children, with the following changes: Adults saw written instructions on top of the page instead of hearing pre-recorded audio files. The three objects were positioned at the bottom of the page (see Figure S1). Furthermore, adults were tested in two additional categories with the following objects:

- instruments: drum, flute, guitar, piano, trumpet, violin, xylophone
- furniture: bed, chair, table, closet, drawer, sofa, lamp, stool

As a consequence, each adult participant received a total number of six trials, one per category, in a randomized order.

Results

To evaluate whether adults selected the object from the same category above chance (33% correct), we aggregated responses across test trials and ran one sample Bayesian t-tests. We found substantial evidence for this hypothesis (mean proportion correct = 0.67, $BF_{10} = 34.93$). Figure S2 shows the corresponding posterior distribution in comparison to the results with children reported in the main manuscript. Like children, adults tracked the topic guiding the interaction with the speaker and used it to make inferences about the referent of an ambiguous pronoun.

Figure S1. Left: Screenshot from the experimental setup for adults.

Supplementary Experiment

In this study, we varied the number of training trials before the test event. Procedures and analysis were pre-registered at https://osf.io/x2k4p/?view_only=d3440e61d88e42c892e3294a6072cc05. 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

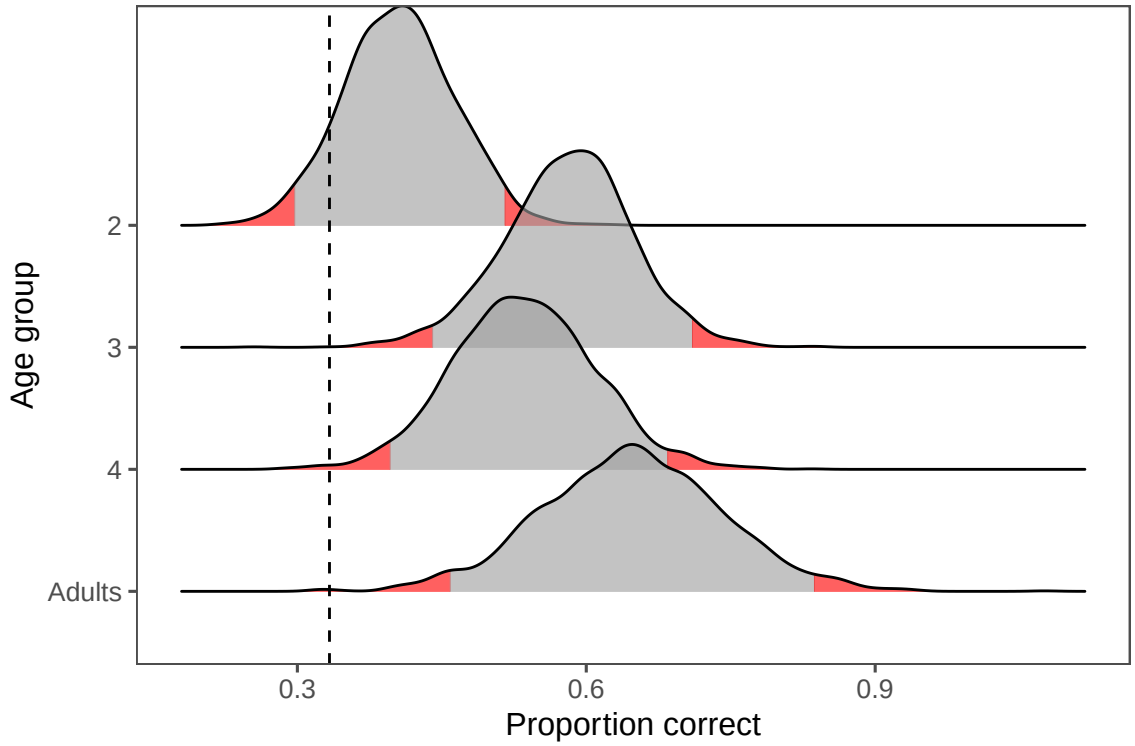

Figure S2. Posterior probability distribution for the mean for each age group and for adults based on one sample Bayesian t-test. Grey regions indicate 95% credible intervals for each age group.

of training trials that preceded the test. *Low input* trials had one training trial while *high input* trials had six training trials. 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 positive, but not reliably different from zero (β = 1.09, 95% CI = -0.71 - 3.08) suggesting a slight increase in performance with age. Figure S3 shows that, in contrast to study 1, 3-year-olds had difficulties with this version of the task.

Table S2
Model comparison for Experiment 2

Model	WAIC	SE	weight
correct ~ age + RE	178.93	9.40	0.61
correct ~ age + condition + RE	180.51	9.90	0.28
correct ~ age * condition + RE	182.34	10.46	0.11

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

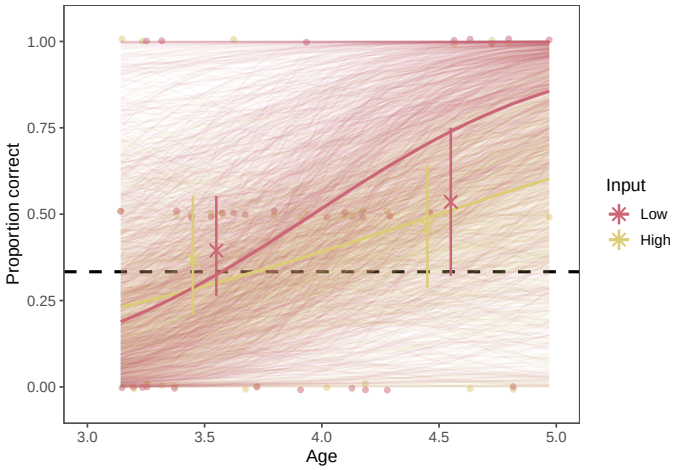

Figure S3. Correct responses in supplementary experiment by age and condition. Transparent dots show aggregated data from individual participants. Blue and red crosses show mean within age bin based on aggregated data with 95% CI based on non-parametric bootstrap. Colored lines show the mean of the posterior 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.

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 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 trials 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.

Analysis. We compared three models. The interaction model included an interaction between age and category, the main effects model included age and category as main effects and the baseline mode only included age as a predictor. Models were compared in the same way as in previous experiments. Furthermore, to investigate the relative difficulty of each target category, we compare the posterior distributions of the parameter estimate for each category.

Results. Figure S4A shows that performance differed widely across categories. This visual impression was corroborated in the model comparison. Table S3 shows that the main effects model with category as a predictor was clearly favored. For comparisons between categories we set animals as the reference category. Figure S4B and S5 show that children chose the object from the same category most often when fruits was the target category. Vehicles and animals resulted in intermediate levels of performance while clothes was the most difficult category.

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

Table S3
Model comparison for exploratory analysis on category differences

Model	WAIC	SE	weight
correct ~ age + category + RE	626.56	12.09	0.85
correct ~ age * category + RE	630.13	12.93	0.14
correct ~ age + RE	637.92	8.55	0.00

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

either because some category members might be less familiar to children and/or because children are less familiar with the category itself. The general discussion section in the main manuscript offers a discussion of how these results relate to the other studies.

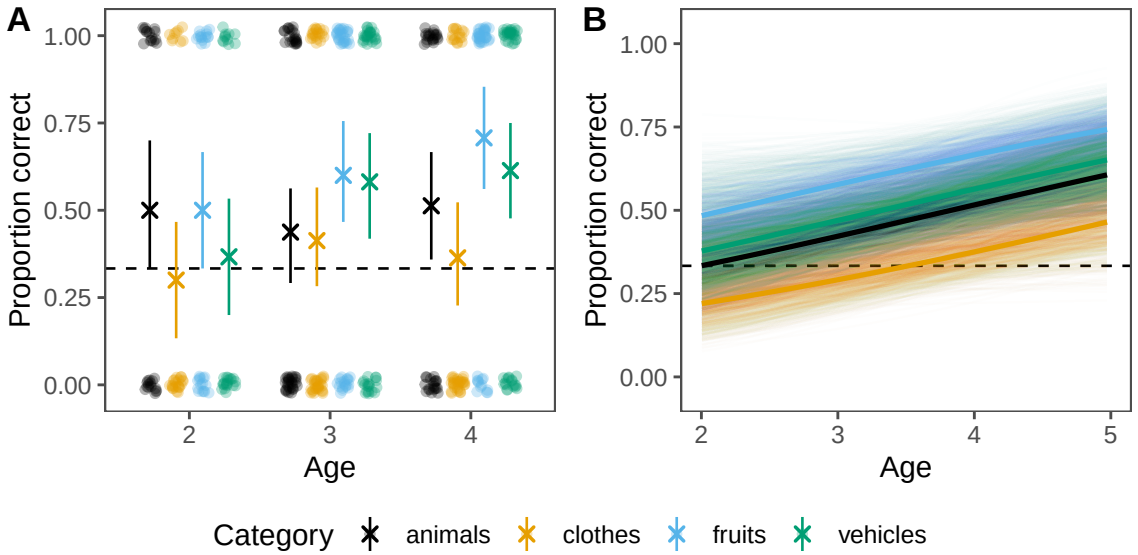

Figure S4. (A) Data plotted by age bin and category. Transparent dots show data from individual participants. Colored crosses show mean within age bin and category with 95% confidence intervals based on non-parametric bootstrap. (B) Correct responses for age continously. Colored lines show the mean of 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.

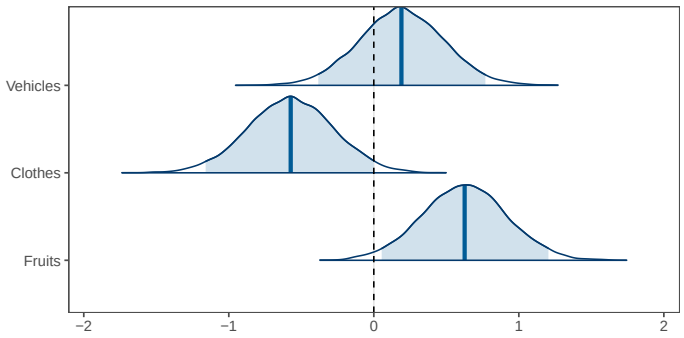

Figure S5. Posterior distribution of model estimates for each category with animals as the reference category. Thick blue lines show mean and shaded regions show 95% credible intervals.

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. https://doi.org/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. https://doi.org/10.32614/RJ-2018-017

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

peerj.preprints.3188v1

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

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., & Gelman, A. (2019). Visualization in bayesian workflow. *J. R. Stat. Soc. A*, *182*(2), 389–402. https://doi.org/10.1111/rssa.12378

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

Wickham, H., & Henry, L. (2019). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

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

Wilke, C. O. (2020). *Ggridges: Ridgeline plots in 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggridges

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