Baguley, Thom; Cahoon, Abbie; Lazareanu, Daniela; Thives Mello, Arthur; Zaneva, Mirela

Working Paper

A comment on "Why Do Children Think Words Are Mutually Exclusive?" (Brody et al., 2024)

I4R Discussion Paper Series, No. 273

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Baguley, Thom; Cahoon, Abbie; Lazareanu, Daniela; Thives Mello, Arthur; Zaneva, Mirela (2025) : A comment on "Why Do Children Think Words Are Mutually Exclusive?" (Brody et al., 2024), I4R Discussion Paper Series, No. 273, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/333868

INSTITUTE for REPLICATION

No. 273
I4R DISCUSSION PAPER SERIES

A comment on "Why Do Children Think Words Are Mutually Exclusive?" (Brody et al., 2024)

Thom Baguley
Abbie Cahoon
Daniela Lazareanu
Arthur Thives Mello
Mirela Zaneva

December 2025

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 273

A comment on "Why Do Children Think Words Are Mutually Exclusive?" (Brody et al., 2024)

Thom Baguley[1], Abbie Cahoon[2], Daniela Lazareanu[3], Arthur Thives Mello[4], Mirela Zaneva[5]

[1]Nottingham Trent University, Nottingham/Great Britain
[2]Ulster University, Belfast/Great Britain
[3]University of Reading/Great Britain
[4]Brunel University of London/Great Britain
[5]University of Oxford/Great Britain

DECEMBER 2025

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

A comment on "Why Do Children Think Words Are Mutually Exclusive?" (Brody et al., 2024)

Baguley, Thom (Nottingham Trent University; thomas.baguley@ntu.ac.uk)
Cahoon, Abbie (Ulster University; a.cahoon@ulster.ac.uk)
Lazareanu, Daniela (University of Reading, Henley Business School;
d.m.lazareanu@pgr.reading.ac.uk; dani.lazareanu@gmail.com)
Thives Mello, Arthur (Brunel University of London; Arthur.ThivesMello@brunel.ac.uk)
Zaneva, Mirela (Christ Church, University of Oxford; mirela.zaneva@chch.ox.ac.uk)
(Authors ordered in alphabetical order by last name)

The authors report no conflicts of interest. No financial support was received for this work.

Abstract

An early emerging understanding of information structure in spoken language is demonstrated in Brody et al.'s (2024) study, contrary to the general assumption that young children innately assume new and familiar words to be mutually exclusive. Across 3 experiments, learners up to 2 years of age (N = 106) showed mutual exclusivity if the novel words were spoken with focus, an information-structural marker of contrast. We successfully computationally reproduced the reported results, ran additional robustness tests, and considered potential covariates that were recorded but not used in the original paper (i.e. age and gender). We also assess potentially disproportionate effects of influential participants via Cook's distance, and refit a model where their data are removed. Further, we consider alternative specifications of random effects (e.g., maximal models) to address convergence issues and obtain interval estimates of experimental conditions' means that account for random effects. Finally, we examine whether different optimizers address model convergence issues. Across our replication tests, we find that the original results are robust and key estimates remain significant. Of note, our influence analysis and Bayesian analysis showed stronger effects than those originally reported, likely attributable to inattentive/noisy responding to some trials. Overall, our replication efforts should provide increased confidence in the original effects.

1. Introduction

Brody, Feiman and Aravind (2024) question the traditional views of mutual exclusivity in early word learning, which take innate inductive bias as the main driver. Instead, the authors examine young children's understanding of information structure (prosody) and the way this is transmitted (prosodic emphasis or focus). The authors focus on the contrastive information naturally contained within spoken language, testing whether this evokes mutual exclusivity or not. They use a variation of the standard mutual exclusivity paradigm involving a novel object choice task, whereby two novel objects are introduced by a central animated character with *focus* versus with *givenness*. Across 3 Experiments, results show that children (N = 106 from across the United States, age range: 24-35 months) only exhibited a mutual exclusivity bias when the novel noun was focused. When it was given (not emphasized), children tended to interpret it (point to) as referring to a previously labeled object. Although prior theories tie mutual exclusivity to word novelty, the present account provides evidence for the important role of information-structural markers, rather than lexical novelty itself, in disambiguating meanings during early language acquisition.

In the present report, collaboratively prepared as part of the Replication Games, an initiative supported by the Institute for Replication and Psychological Science, we investigate whether Brody and colleagues' results are computationally reproducible and carry out a number of robustness tests:

1) considering the addition of age and gender as covariates,
2) assessing sample sensitivity to influential participants and refitting a model without such participants,
3) using different optimization algorithms,
4) re-expressing the categorical predictors under different contrast schemes
5) considering different specifications of the random structures (e.g. maximal models), comparing point and interval estimates between those reported in the paper (derived from raw data), those from the models used for

significance tests in the paper (but not used to derive estimates) and Bayesian models with the appropriate maximal random effects.

We successfully reproduce the results and note only a minor deviation from optimal practices, where two different models were stored under the same name ("m3"). We would like to acknowledge that the original study was first reproduced by the Psychological Science STAR team, who confirmed the computational reproducibility of the results.

Turning to our sensitivity analyses (robustness reproduction), we find that including additional covariates (age, gender) in the models does not change the pattern of significance for the main results. The point estimate for *condition* changes slightly (original $\beta = -12.712$, se = 4.490, $p = 0.005$ vs $\beta = -11.850$, se = 4.934, $p = 0.0163$ for model with age, vs $\beta = -10.296$, se = 3.356, $p = -.0022$ for model with gender, vs $\beta = -10.188$, se = 3.285, $p = 0.0019$ for model with age + gender; see Table 2). Neither age nor gender were significant factors in our analyses. Next, we used Cook's distance to examine whether there were any influential participants. 6 participants (associated with 33 trials) were influential (according to a common rule-of-thumb threshold of 4/n), thus we removed them in a follow-up test. The key estimate for condition was robust, remained significant, and became stronger (original model $\beta = -12.712$, se = 4.490, $p = 0.005$ vs model with exclusion of influential participants $\beta = -18.864$, se = 3.406, $p < 0.0001$).

In a separate set of tests, we refitted the models using multiple optimizers, including BOBYQA (original), Nelder–Mead, NLMINB, and two NLopt-based variants (NLOPT-BOBYQA and NLOPT-Nelder–Mead). Across all optimizers, estimates remained unchanged among the models that converged. We further examined robustness to contrast specification by applying reverse-treatment and sum coding, confirming the results are robust to contrast specification. Finally, point and interval estimates for the percentage of new objects selected (the key outcome measure) were derived from raw data whereas significance tests were obtained from a mixed effects generalized linear model with a binomial link.

2. Computational Reproducibility

We summarize the computational reproducibility of the study in regards to the provided data and code in Table 1. The authors provided data in the form of 3 csv files (one for each experiment), alongside a Rmd file with code and the generated html output. The Rmd file contained the complete code, including data wrangling, analysis, and figure generation. The Rmd file ran without errors and produced the same output file as shared by the authors. We checked this against the paper, and the results matched. In other words, we successfully reproduced Experiments 1, 2, 3 (all reported tests). We note that the data provided contains only the included participants, though the authors transparently describe the numbers of excluded participants with reasons in text. The authors have also labeled their data files as *'cleandata*' (e.g. foxy_exp1_cleandata_wide2.csv). This is why we treat this as 'cleaned' and not 'raw' data. Similarly, although the provided code includes relevant data wrangling (e.g. wide to long format transitions), it does not contain all relevant cleaning code – for instance, it does not check for missing trials or other exclusion criteria.

	Fully	**Partial**	**No**	**Notes**
Raw data provided			x	Cleaned data provided. Sufficient for replication of results.
Cleaning code provided		x		Code contains some relevant data wrangling but not all (e.g. exclusions)
Analysis data provided	x			
Analysis code provided	x			
Reproducible from raw data			x	Raw data not provided
Reproducible from analysis data	x			

Table 1. Summary table for the provided data and code with regard to reproducibility.

As a minor note, two different models were saved as ‘m3’ in the original code. This was not an issue in Markdown (which prints outputs in sequence, thus model outputs are visible clearly for each model) but could be an issue if someone later wanted to call the earlier model, which had been overwritten with the later instance of ‘m3’. We recommend that the second model is saved with a different name, e.g. ‘m3_e3’ for anyone reusing these analyses.

2.1 Discrepancies Between Pre-analysis Plan and Article

The authors have pre-registered each of the three experiments, including pre-analysis plan and description of the data collection, procedure, statistical models, inference criteria, and participant exclusion criteria.

We find no deviations from the pre-analysis plan. We note that there was a non-completion criterion, used for potential exclusion of participants, that was defined as “Does not complete the majority of the experimental items (at least 3/6 usable trials for inclusion).” We did not see code for this included in the shared Rmd file. Nevertheless, data shared included only participants with at least 3/6 trials. In other words, we believe the authors likely checked for this and excluded participants earlier.

We also note in passing that the authors preregistered a significance criterion of “p = 0.05”, which we take to mean $p < 0.05$ or perhaps even ≤ 0.05. This criterion is inconsistently applied in the paper in one instance where the result is described as “*albeit only marginally significant ($\beta = 4.06$, $p = .077$)*” (Brody et al., 2024, p. 1321).

3. Robustness Reproductions

In this section, we report a series of robustness tests, examining: 1) additional covariates, 2) sample sensitivity, 3) different optimizers, 4) contrast coding, 5) different specifications of the random structures (e.g. maximal models).

3.1 Robustness test: Additional covariates (age and gender)

Here, we focus on the mixed effect logistic regression model for Experiment 1 (with glmer package in R) and consider the addition of covariates. The data contains variables for

participants' age in months (variable: "AgeMonths") and gender (variable: "Gender"). Given the potential rapid and heterogenous developmental changes in early childhood (Lewis et al., 2020; Bergelson & Swingley, 2012), we consider it might be sensible to include age as a covariate. We did not anticipate gender differences in very young children but consider the addition of this variable as a robustness test.

As a note on data cleaning, we found that some values for 'Gender' had an additional space. We ran a simple check for unique values in R: unique(exp1.wide$Gender). We noticed the output showed two different entries for male participants: [1] "F" "M" "M ". We cleaned this by treating values as "M" and "M " as both "M" (i.e., both referring to male participants).

We first examine one of the focal results for Experiment 1, namely for the rate of mutual exclusivity inferences in the focus group: β = -12.712, *p* = 0.005 (p. 1320 of the original manuscript). In the authors' code notation, this model essentially asks whether two conditions (focus vs givenness) differ from each other. Condition is dummy-coded with the focus condition as the reference level.

In Table 2, we report three further robustness models. These are analogous to the original (base) model but with added covariates: 1) base model + age, 2) base model + gender, 3) base model + age + gender. The estimates for age and gender are not significant in any of these models. The estimate for Condition remains significant in all models, with relatively small changes to the exact estimate value (original β = -12.712 vs β = -11.850 for model with age, vs β = -10.296 for model with gender, vs β = -10.188 for model with age + gender). In other words, we find this estimate to be robust to the inclusion of covariates.

Model	Intercept (SE) *p-value*	Condition (*SE*) *p-value*	Age (*SE*) *p-value*	Gender (*SE*) *p-value*
Base (Original)	7.289*** (2.059) *p = .0004*	-12.712** (4.490) *p = .005*	-	-

Base + Age	1.418 (9.734) *p = .8842*	-11.850* (4.934) *p = .0163*	0.183 (0.300) *p = .5425*	-
Base + Gender	5.202** (1.967) *p = .0082*	-10.296** (3.356) *p = .0022*	-	2.342 (1.713) *p = .1717*
Base + Age + Gender	0.243* (8.7402) *p = .9778*	-10.188* (3.285) *p = .0019*	0.165 (0.289) *p = .5686*	2.233 (1.703) *p = .1897*

Table 2. Robustness test considering the addition of covariates. We report the original model (which we refer to as the base model here). We then report three robustness tests, where we add age (in months) as a covariate to the base mode, then gender, then both age and gender. *** p < 0.001, ** p < 0.01, * p < 0.05.

3.2 Robustness test: Sample sensitivity with Cook's distance

To assess the impact of potentially influential participants on the mixed-effects model in Experiment 1 (model labelled 'm1' in the authors' code, referred to as 'Base' model in the previous section 3.1 in our report), we looked at Cook's distance (influence.ME package in R). We computed influence measures at the participant level (i.e. used "Participant" as the grouping variable). We looked at two rule-of-thumb thresholds (Bollen & Jackman, 1985; Hair et al., 1998) in order to classify participants as influential, specifically 4/n and 4/(n-k-1), where n is the number of participants and k is the number of fixed effects in the model. In Figure 1, we show a scatterplot of Cook's distance values. 6 participants were above both of these thresholds (all other participants were below the more liberal threshold). These 6 participants completed a total of 33 trials.

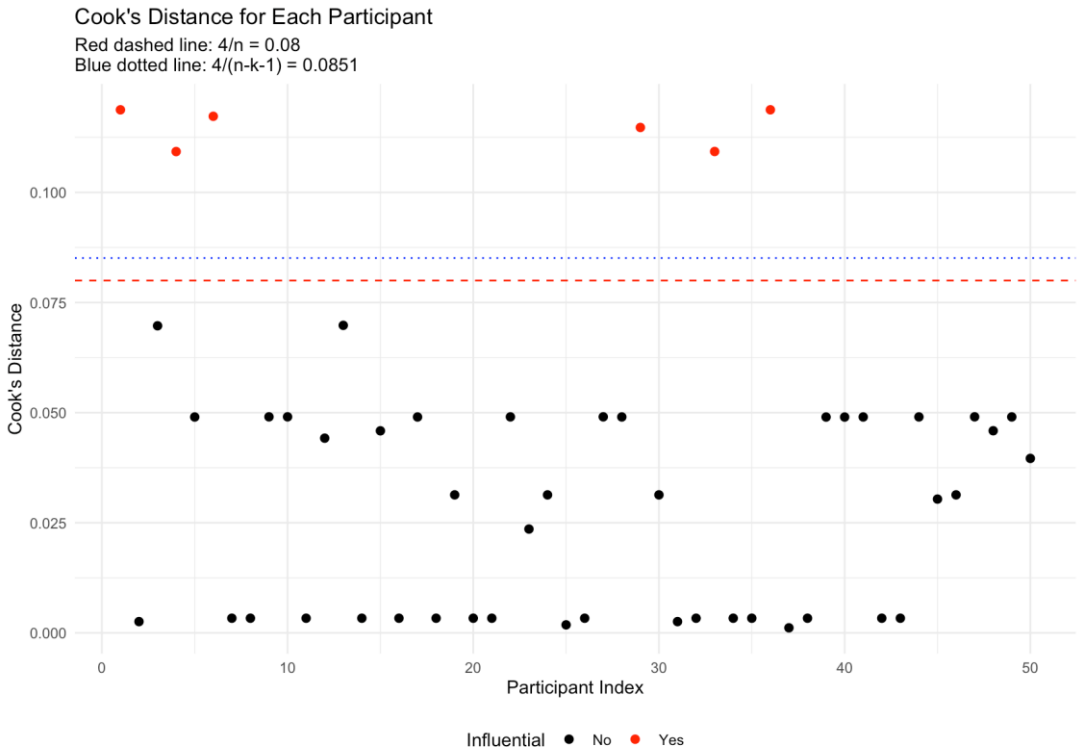

Figure 1. Scatterplot for Cook's distance. 6 participants (red filled dots) are considered influential, as they are above both considered thresholds. The two thresholds are presented with red and blue dotted lines.

Next, we refit the base model without the influential cases (Table 3). In short, following this exercise, the coefficient for Condition remains significant and becomes more pronounced (original model β = -12.712 vs. model with exclusion of influential participants β = -18.864). The result is robust, and in our replication, now stronger. We speculate that the influential participants might have been (somewhat) inattentive responders.

Model	Intercept (*SE*) *p-value*	Condition (*SE*) *p-value*
Base (Original)	7.289*** (2.059) *p = .0004*	-12.712** (4.490) *p = .005*
Base – 6 influential participants removed	9.819*** (2.092) *p < .0001*	-18.864*** (3.406) *p < .0001*

Table 3. Robustness test considering sample sensitivity. We report the original model (which we refer to as the base model here). We then report the same model, re-ran with the exclusion of 6 influential participants as determined by Cook's distance. *** p < 0.001, ** p < 0.01, * p < 0.05.

3.3 Robustness test: Different optimizers

Fitting each model with different optimization algorithms produced indistinguishable fixed-effect estimates (Table 4), when considering models that converged, indicating that the original authors' results are robust to optimizer choice.

Model	Optimizer	Term	Converged	Singular	Estimate	Standard error	Z value	P value	Lower bound 95% CI	Upper bound 95% CI
m1	bobyqa*	ConditionG	TRUE	FALSE	-12.71161619	4.48963780	-2.83132329	0.00463558	-21.51130624	-3.91192605
m1	Nelder_Mead	ConditionG	TRUE	FALSE	-12.71162435	4.48945123	-2.83144279	0.00463385	-21.51094877	-3.91229991
m1	nlminbwrap	ConditionG	TRUE	FALSE	-12.71162556	4.48975026	-2.83125447	0.00463658	-21.51153609	-3.91171503
m1	nloptwrap.NLOPT_LN_NELDERMEAD	ConditionG	TRUE	FALSE	-12.71162801	4.48764828	-2.83258116	0.00461738	-21.50741866	-3.91583736
m1	nloptwrap.NLOPT_LN_BOBYQA	ConditionG	TRUE	FALSE	-12.71165201	4.48870963	-2.83191675	0.00462698	-21.50952293	-3.91378112
m2	bobyqa*	(Intercept)	TRUE	FALSE	-2.903654321	1.15426746	-2.51558188	0.01188360	-5.166018548	-0.641290094
m2	Nelder_Mead	(Intercept)	TRUE	FALSE	-2.971680114	1.30603063	-2.27535253	0.02288479	-5.531500153	-0.411860068
m2	nlminbwrap	(Intercept)	TRUE	FALSE	-2.974551352	1.31411851	-2.26353561	0.02360281	-5.550223637	-0.398879072
m2	nloptwrap.NLOPT_LN_NELDERMEAD	(Intercept)	TRUE	FALSE	-2.974495231	1.31402650	-2.26364934	0.02359569	-5.549987179	-0.399003282
m2	nloptwrap.NLOPT_LN_BOBYQA	(Intercept)	TRUE	FALSE	-2.974515456	1.31412593	-2.26349346	0.02360528	-5.550202284	-0.398828628
m3	bobyqa*	ExperimentExp2	TRUE	FALSE	0.794416019	1.47346772	0.53914721	0.58978528	-2.093580728	3.682412766
m3	Nelder_Mead	ExperimentExp2	TRUE	FALSE	0.793612927	1.47188468	0.53918145	0.58976165	-2.091281064	3.678506918
m3	nlminbwrap	ExperimentExp2	TRUE	FALSE	0.794413792	1.47354368	0.53911791	0.58980549	-2.093731827	3.682559408
m3	nloptwrap.NLOPT_LN_NELDERMEAD	ExperimentExp2	TRUE	FALSE	0.794451001	1.47353832	0.53914512	0.58978672	-2.093684119	3.682586121

m3	nloptwrap.NLOPT_LN_BOBYQA	ExperimentExp2	TRUE	FALSE	0.79464938	1.47359031	0.53926072	0.58970696	-2.093587636	3.68288639
m4	bobyqa*	(Intercept)	TRUE	FALSE	4.06016676	2.29452842	1.76949943	0.07681056	-0.437108951	8.55744247
m4	Nelder_Mead	(Intercept)	FALSE	TRUE	3.863232079	1.93496549	1.99653796	0.04587539	0.070699704	7.655764454
m4	nlminbwrap	(Intercept)	TRUE	FALSE	4.06016698	2.29455324	1.76948039	0.07681373	-0.437157379	8.55749133
m4	nloptwrap.NLOPT_LN_NELDERMEAD	(Intercept)	TRUE	FALSE	4.06034040	2.29481436	1.76935462	0.07683471	-0.437495747	8.55817655
m4	nloptwrap.NLOPT_LN_BOBYQA	(Intercept)	FALSE	TRUE	3.863568266	1.93547739	1.99618362	0.04591394	0.070032581	7.657103952
m5	bobyqa*	(Intercept)	TRUE	FALSE	3.863240352	1.93498834	1.99651866	0.04587749	0.070663198	7.655817506
m5	Nelder_Mead	(Intercept)	TRUE	FALSE	3.863232386	1.93499339	1.99650934	0.04587851	0.07064534	7.655819432
m5	nlminbwrap	(Intercept)	TRUE	FALSE	3.863240635	1.93502130	1.99648481	0.04588117	0.070598886	7.655882385
m5	nloptwrap.NLOPT_LN_NELDERMEAD	(Intercept)	TRUE	FALSE	3.862931254	1.93457318	1.99678734	0.04584828	0.071167814	7.654694694
m5	nloptwrap.NLOPT_LN_BOBYQA	(Intercept)	TRUE	FALSE	3.863208711	1.93497390	1.99651721	0.04587765	0.070659853	7.655757577
m6	bobyqa*	ExperimentExp3	TRUE	FALSE	7.804606117	2.16221036	3.60954986	0.00030672	3.566673811	12.04253842
m6	Nelder_Mead	ExperimentExp3	TRUE	FALSE	7.802391139	2.15784905	3.61581878	0.0002994	3.573006995	12.03177528
m6	nlminbwrap	ExperimentExp3	TRUE	FALSE	7.804626229	2.16212569	3.60970051	0.00030655	3.566859873	12.04239259
m6	nloptwrap.NLOPT_LN_NELDERMEAD	ExperimentExp3	TRUE	FALSE	7.805357838	2.16248106	3.60944562	0.00030685	3.566894943	12.04382073
m6	nloptwrap.NLOPT_LN_BOBYQA	ExperimentExp3	TRUE	FALSE	7.804437156	2.16213011	3.60960568	0.00030666	3.566662126	12.04221219
m7	bobyqa*	ExperimentExp3	TRUE	FALSE	-0.540536395	1.43550048	-0.376549085	0.70650873	-3.354117349	2.27304455
m7	Nelder_Mead	ExperimentExp3	FALSE	TRUE	-0.536960465	0.00121629	-441.4718866	0	-0.539344405	-0.535476525
m7	nlminbwrap	ExperimentExp3	FALSE	TRUE	-0.542717589	1.43129743	-0.379178761	0.70455512	-3.348060566	2.26262538
m7	nloptwrap.NLOPT_LN_NELDERMEAD	ExperimentExp3	TRUE	FALSE	-0.535816842	1.44548528	-0.370683013	0.71087364	-3.368968003	2.29733432

Table 4. Fixed effect estimates for each model using five different optimization algorithms and their associated convergence status. *Note*. *Indicates the optimizer used by the original study authors (bobyqa in all models). M1: Tests whether the two conditions differ. Condition is dummy-coded with the Focus condition as reference. M2: Tests whether the rate of selecting the new referent differs from chance. M3: Tests whether the Given conditions differ between Experiment 1 and Experiment 2. Experiment is dummy-coded with Experiment 1 (Given) as reference. M4: Tests whether the rate of selecting the new referent differs from chance in Experiment 3. M5: Exploratory model testing whether the

rate of selecting the new referent differs from chance without an item intercept. M6: Tests whether the rate of selecting the new object differs between Experiment 2 and Experiment 3. M7: Tests whether the F-conditions differ between Experiment 1 (Focus) and Experiment 3. Experiment is dummy-coded with Experiment 1 (Focus) as reference.

3.4 Robustness test: Contrast coding

Re-expressing the categorical predictors under different contrast schemes (treatment and sum coding) did not change the estimated pairwise log-odds contrasts (Table 5), confirming that the results are robust to different coding schemes.

Model	Coding	Contrast	Estimate (link)	Standard error (link)	Lower bound 95% CI	Upper bound 95% CI
m1	treatment_first*	F - G	12.7116161	4.48963781	3.91208772	21.5111445
m1	treatment_reversed	G - F	-12.711613	4.48521487	-21.502473	-3.9207537
m1	sum_coding	F - G	12.7116194	4.48919876	3.91295146	21.5102872
m3	treatment_first*	Exp1 - Exp2	-0.794416	1.47346773	-3.6823597	2.09352766
m3	treatment_reversed	Exp2 - Exp1	0.79441399	1.47350846	-2.0936095	3.68243749
m3	sum_coding	Exp1 - Exp2	-0.7944144	1.473537	-3.6824939	2.09366504
m6	treatment_first*	Exp2 - Exp3	-7.8046061	2.16221036	-12.042461	-3.5667517
m6	treatment_reversed	Exp3 - Exp2	7.80460708	2.16222682	3.56672039	12.0424938
m6	sum_coding	Exp2 - Exp3	-7.8046079	2.16222106	-12.042483	-3.5667324
m7	treatment_first*	Exp1 - Exp3	0.54053639	1.43550049	-2.2729929	3.35406565
m7	treatment_reversed	Exp3 - Exp1	-0.5405346	1.43547154	-3.3540071	2.27293792
m7	sum_coding	Exp1 - Exp3	0.54053488	1.43548538	-2.2729648	3.35403453

Table 5. Pairwise log-odds contrasts between conditions for each model under three contrast-coding schemes (treatment, reversed treatment, and sum coding). *Note*. *Indicates the coding used by the original study authors (treatment first in all models). M1: Tests whether the two conditions differ. Condition is dummy-coded with the Focus condition as reference. M3: Tests whether the Given conditions differ between Experiment 1 and Experiment 2. Experiment is dummy-coded with Experiment 1

(Given) as reference. M6: Tests whether the rate of selecting the new object differs between Experiment 2 and Experiment 3. M7: Tests whether the F-conditions differ between Experiment 1 (Focus) and Experiment 3. Experiment is dummy-coded with Experiment 1 (Focus) as reference.

3.5 Robustness test: Robustness to different random effects specifications

The original authors first fitted models with the structure y ~ 1 + (1|item) + (1participant) or if a categorical predictor was present y ~ 1 + Factor + (1|item) + (1participant). These were fitted with the R package lme4 (Bates et al., 2015). However, some of these fits produced warnings in relation to the item effect and models with simpler random effects structures (e.g., excluding item effects) were reported in some cases. This is unsatisfactory for several reasons. First, deriving estimates and tests from different analyses can lead to inconsistencies. Second, the estimates from raw data handle missing trials in a suboptimal way (assuming trials are MCAR; missing completely at random) relative to a mixed effect model (which treats missing trials as MAR; missing at random). Third, the raw data approach ignores the random effects.

It is not clear why estimates were derived from raw data rather than the model used for inference. This could be due to warnings in relation to convergence for some models; alternatively, due to interval estimates from these models relying on asymptotics - whether profiling likelihood or a Wald style (*z*) approximation is used. It is well known that accurate interval estimates and tests for mixed effects models are not straight-forward to obtain. Bolker et al. (2025) suggest that methods for obtaining inferences from single parameters are (from worst to best) Wald *z* tests, Wald *t* tests, Likelihood ratio tests and Markov Chain Monte Carlo (MCMC) or parametric bootstrap interval estimates.

For this robustness check we refitted all the models as maximal models using Bayesian mixed effects logistic regression models with the Bernoulli family in the brms package in R (Bürkner, 2017). This uses MCMC estimation to fit the models and is considered one of the gold standard approaches to obtaining accurate inferences. Under this approach missing trials are assumed to be MAR. We used default (weakly informative priors in

brms) to obtain interval estimates: specifically Highest Posterior Density (HPD) intervals. With this prior specification it is expected that the interval estimates should closely match those from a frequentist mixed effects model. The maximal model specification broadly follows recommendations from Barr et al. (2013). The aim is to capture the structure of the data with the effects being modeled. The key models either take the form of a fixed intercept with random effects or a fixed intercept plus factor (either Condition with levels focus vs. given or Experiment with two levels representing experiments being compared). We therefore fit the following random effect structure for intercept only models:

$$y \sim 1 + (1|item) + (1|participant)$$

In these models the focus is on whether the choice of the new item was above chance (50%) for the focus condition or below chance for the given condition. For the models with a fixed factor we fit the structure:

$$y \sim 1 + (1|item) + (1 + Factor||participant)$$

As the factor (Condition or Experiment) always represents a comparison between two groups of participants the common maximal structure (1 + Factor|participant) is inappropriate because it includes a correlation term between the random intercept and dummy coded Factor. As these are independent groups the expectation is that the effects are uncorrelated, but (1 + Factor||participant) in effect fits different random intercepts with distinct variances for each group. This is potentially important because the variances do differ between conditions and experiments (sometimes considerably).

While we fitted maximal Bayesian versions for each of the reported models the overall pattern of results can be summarized by comparing just the three key models M1, M2 and M3 representing the primary analysis for each of the three experiments. Table 6 compare the point and interval estimates for each of the experiments. In each case the outcome is the percent of new objects selected by the participant. These are: i)

estimates derived from raw data (that were reported in the original paper), ii) estimates derived from the models used to derive the hypothesis tests in the original paper (for which point and interval estimates were not reported), and iii) the estimates from the maximal Bayesian models fitted as a robustness check.

3.6. Robustness test comparing point and interval estimates derived from raw data, the reported mixed effect model and a maximal Bayesian model

From Table 6 several clear patterns emerge. The point estimates from the raw data differ from those for the lme4 models used to derive hypothesis tests for each experiment, while the models used to derive hypothesis tests and the maximal Bayesian model estimates are more similar, but not identical. The raw data interval estimates are symmetrical and narrower than the Wald estimates or the Bayesian HPD intervals. Overall, the Bayesian estimates provide stronger support for the hypotheses than either the raw data or models used for hypothesis testing by Brody and colleagues.

	Experiment	***E1***	***E2***	***E3***
	Model Label	*M1*	*M2*	*M3*
		Focus		*Focus*
Raw data	Percentage new	89.4%	—	86.7%
	95% interval estimate	84.3, 94.3	—	79.8, 93.3
Reported model	Percentage new	99.3%	—	98.3%
	95% interval estimate	96.3, 100.0	—	39.2, 100.0
Bayesian model	Percentage new	99.2%	—	96.0%
	95% interval estimate	95.1, 100.0	—	83.8, 100.0
		Given	*Given*	
Raw data	Percentage new	24.3%	29.4%	—
	95% interval estimate	17.3, 31.2	23.6, 35.4	—
Reported model	Percentage new	0.4%	5.2%	—

	95% interval estimate	< 0.5, 47.8	0.6, 34.5	—
Bayesian model	Percentage new	3.5%	9.9%	—
	95% interval estimate	< 0.5, 18.9	< 0.5, 33.0	—

Table 6. Robustness test comparing point and interval estimates derived from raw data, the reported mixed effect model and a maximal Bayesian model. *Note*. Raw data estimates are from Brody *et al*. (2024). Raw data interval estimates are from bootstrapping raw data. Reported model estimates are from the labeled model in the Brody *et al*. R script fitted with lme4. Corresponding interval estimates are Wald style normal approximation. Bayesian model estimates are from a fitted model in brms with default priors in R with maximal random effects as appropriate to the model (see full explanation in text). Interval estimates are HPD intervals from brms.

The most striking feature is the difference in the point estimates for the raw data and the estimates from either of the mixed effects models. This discrepancy requires some explanation. In a mixed effect model the estimates exhibit shrinkage. In essence this shifts (“shrinks”) estimates towards mean values and reduces the impact of atypical responses (those with large residuals). Shrinkage makes the model slightly worse at describing the observed data but improves out of sample generalization. The degree of discrepancy in this study is however larger than typically seen. This is in part because of the small sample size but largely because of the nature of the data and sample. With children in this age range (2 - 3 years) some are likely inattentive and essentially responding at random on one or more trials. The exclusion criteria removed participants fewer than 3 usable trials out of six. Figure 2 in Brody et al. (2024, p. 1320) indicates that typical children selected the new object near ceiling on focus trials and near floor on given trials (with the modal response being 100% or 0% in many cases). Thus the model estimates are shrunk towards the more consistent and (presumably) less distracted or inattentive children. The analysis of raw data also exacerbates any differences because of missing trials. Averaging percentages per participants weights trials from children with missing data more than those with complete cases (who are also likely the less distracted participants). This is also consistent with the influence

analysis reported earlier (where removing the higher influence data points strengthened the effects).

(Shrinkage also occurs for the items, but item variance was negligible, and it is unlikely to have had much impact on the estimates in any of the models. Estimation issues in the original studies were in fact largely singularity warnings when the item variance was estimated as zero or negative.

In terms of inferences, the interval estimates from the maximal Bayesian model are more accurate than the CIs derived from the raw data and or the Wald CIs for reasons outlined above. They provide somewhat stronger support for the main findings of Brody et al. (2024) than the reported interval estimates and tests. This is most apparent for M3 in Table 6. The reported test of this effect in Experiment 3 is a Wald test of the log odds of the intercept vs. zero (i.e., selecting the new object above the 50% chance). This is non-significant with $\beta = 4.06$, $z = 1.769$, $p = .0768$ (and a corresponding interval estimates that includes 50%). A more accurate likelihood ratio test for the same model gives $\chi^2(1) = 15.73$, $p = .0073$. This is also supported by the Bayesian HPD interval of [83.8, 100] which comfortably excludes 50%. We can also derive a Bayesian approximation to a *p* value for this test. A one-sided frequentist *p* value can be considered an approximation to the Bayesian posterior probability of an effect in the opposite direction to that observed (Casella & Berger, 1987). For this test, this probability is .001 and thus a Bayesian approximation to the two-sided *p* value (for comparison) is $p = .002$.

4. Conclusion

Our replication efforts confirm the computational reproducibility and robustness of Brody et al.'s (2024) original results. We successfully reproduced key effects, and those remained significant across our robustness tests, including the addition of covariates, the removal of influential participants, varying optimizers or contrast coding, and fitting maximal Bayesian models. Notably, in two of our tests (influence and Bayesian analyses) we found support for stronger effects than the ones originally reported by the authors. We

speculate this could be because of slightly inattentive responding on some trials. These stronger effects should not be interpreted as a critique against the authors, who had transparently pre-registered sensible exclusion criteria for inattentive responding. Instead, among attentive children, it should be interpreted as further support for the conditional effect of mutual exclusivity. Altogether, our replication tests give us increased confidence in the results reported by Brody and colleagues.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, W. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences* 109(9), 3253–3258.

Bollen, K. A., & Jackman, R. W. (1985). Regression diagnostics: An expository treatment of outliers and influential cases. *Sociological Methods & Research*, 13(4), 510-542

Bolker, B. *et al*. (2025, 19 July). GLMM FAQ. https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html Retrieved 20 November 2025.

Brody, G., Feiman, R., & Aravind, A. (2024). Why do children think words are mutually exclusive?. *Psychological Science*, *35*(12), 1315-1324.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1-28.

Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association*, *82*(397), 106–111.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis*. Englewood cliff. New jersey, USA, 5(3), 207-2019.

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, 104–191.