A robustness reproduction of Cox et al. (2023)*

Björn S. Siepe†, Matthias Kloft,
Semih C. Aktepe, Daniel W. Heck

November 28, 2024

Abstract

Cox et al. (2023) investigated the acoustic features of infant-directed speech. They used Bayesian meta-analyses to investigate five acoustic features with data from 88 studies. In the present robustness reproduction, we first check if the reported results are reproducible based on the data and code provided by the original authors. We then perform robustness analyses by using different measures of model comparison and by investigating the robustness of the Bayesian sampling approach used by the authors. We find that the main results of Cox et al. (2023) can be reproduced. However, some minor coding errors and unclarity in the results hindered us from reproducing all results in the paper, including those regarding model comparison. In our robustness analyses, we find that alternative ways of model comparison also contradict some of the original results. Further, while the original models were complex and occasionally prone to convergence issues, simplifying them did not show a noticeable impact on the results. Taken together, the reproducibility of the core findings of the paper strengthens its main conclusions, although some results regarding the relevance of moderator variables remain unclear.

[*] Authors: Siepe (corresponding author †): University of Marburg. E-mail: bjoern.siepe@uni-marburg.de. Kloft: University of Marburg. E-mail: kloft@uni-marburg.de. Aktepe: University of Marburg. E-mail: semih.aktepe@uni-marburg.de. Heck: University of Marburg. daniel.heck@uni-marburg.de. All authors declare no conflict of interest. This project is part of the Institute for Replication (I4R) effort to replicate and reproduce articles published in Nature Human Behaviour.

1 Brief summary of the original paper

Cox et al. (2023) performed a Bayesian meta-analysis of acoustic features of infant-directed speech (IDS) compared to adult-directed speech (ADS) using data from 88 unique studies. They investigated five parameters: fundamental frequency (f_0), f_0 variability, vowel space area, articulation rate, and vowel duration. They performed the same analyses for each of these parameters separately. Briefly, they first calculated three-level intercept-only random effect models with language, study, and measures as random effects for each parameter. They then used task, environment, age, and language as predictors/moderators of the effect size in a hierarchical Bayesian model. They compared this full model with alternative model versions without each of the moderator variables based on stacking weights. They mostly interpreted the effect sizes of the model with the highest stacking weight for each of the five parameters in Table 1 of their manuscript.

Given the estimated effect sizes on the five parameters, the authors concluded that IDS differs robustly from ADS and that these findings still hold when accounting for potential publication bias (with one noted exception). They further noted that age influenced f_0, articulation rate, and vowel duration, whereas f_0 variability and vowel space were stable during development. We now turn to our robustness reproduction of these results.

2 Reproducibility

All code and materials for this robustness reproduction are available at the Open Science Framework (`https://osf.io/ukfrc/`). All of the following analyses were conducted in R (version 4.4.2, R Core Team 2024) and run on a Linux machine with an Ubuntu (version 22.04.5) distribution. The complete output of `sessionInfo()` is provided in the online supplement. As in the original paper, we used the `brms` package (Bürkner 2017) for meta-analyses in the probabilistic programming language Stan (Stan Development Team 2023), using `rstan` as a backend

(Stan Development Team 2024). We additionally used the `renv` package to create a reproducible package environment (Ushey and Wickham 2024).

2.1 Computational Reproducibility

We assessed the computational reproducibility of the authors' results by executing all code provided by the authors on the Open Science Framework.[1] We checked if the code ran without error. We additionally checked to see if we could recreate the figures and reproduce the main results in the manuscript. While re-running the code, we checked it for coding errors.

The code for missing value imputation and data cleaning was not available, so we could not check it for correctness or examine the exact settings used. Instead, only the cleaned and imputed data sets were available. Additionally, code or model output was missing for some models that were either mentioned in the paper or referred to somewhere in the code. We fitted these models by writing the code ourselves, which simply involved replacing some predictor variables.

In addition, we found some minor coding errors, such as a misplaced end of a for-loop or an omitted variable in column selection that prevented us from simply running the scripts as provided online, but these errors were easy to correct. We report them in the online supplement.

None of these issues had a drastic impact on the results obtained in our analyses. Rather, they impeded the reproducibility of the results and forced us to redo or revise some of the analyses. We give an overview of these results in the reproduction package in Table 2.

2.2 Model Results

In this section, we report the comparison of point estimates we obtained from re-estimating the models exactly as specified in the original code. We have reproduced the original results reported in Table 1 of Cox et al. (2023) in the appendix (see our

[1] Available at `https://osf.io/hc7me/`.

Table 2). We consider the reproduction of these main results to be successful, as all point estimates and credible intervals were in the same direction and very similar to the original estimates (taking into account the margin of error due to the use of MCMC sampling for parameter estimation). We additionally reproduced Figures 3 to 7 of the original manuscript. At least in our subjective visual assessment, the reproduction of these figures was also successful.

The terminology of “robust predictors” used in Table 1 of the original paper was unclear to us. We assumed that this column referred to predictors whose credible intervals excluded 0 in the best-fitting model since this model was used to report the average effect size. However, this does not seem to be the case. We tried alternative interpretations (see the online supplement), but it remains unclear to us how it was decided whether predictors were considered to be “robust” or not (this is indicated by question marks in Table 2).

For each parameter, the original authors additionally provided model weights to indicate which model provided the best account for the data. In the paper, the authors stated that they used “leave-one-out stacking weights”. This terminology seems unclear, as classical leave-one-out (loo) weights, as implemented in the original code by Cox et al. (2023), are obtained differently than stacking weights (see Yao et al. 2018, for more details). As the authors did not provide the output of the model weights computation and did not provide all relevant model output, we recomputed the weights using their code. Additionally, the original code does not match the description in the paper as model weights referred to in the paper are not computed in the code.[2] Table 3 contains all model weights for all outcomes. Using loo weights, we were unable to reproduce the model weights and ranking of models reported in the manuscript for all outcomes except articulation rate.

[2] For example, for f_0, the authors stated that the model with task, age, and language had a high stacking weight, but this model is not included in the model weight calculation of the authors.

3 Robustness Reproduction

We conducted a robustness reproduction of the main results in the paper. Each of the following reanalyses was applied to all five parameters studied. Note that we do not consider any of these alternative analyses necessarily superior to the analysis choices made by Cox et al. (2023). Due to the computational complexity of the planned statistical analyses, we performed some of the robustness analyses on a subset of the reported analyses.

As preregistered, we defined a reproduction as an effect that showed the same direction as in the original paper and whose 95% Credible Interval excluded 0, or, in the case of a null effect in the original paper, also included 0. In addition, we interpret the potential differences in estimated effect sizes qualitatively.

3.1 Alternative approaches to account for publication bias.

The authors assessed publication bias using the `PublicationBias` R package (Mathur and VanderWeele 2020). This method enables researchers to estimate the minimum severity of publication bias that would be required to attenuate the Credible Interval of the effect size estimate to include zero. However, the authors did not directly estimate pooled effect sizes adjusted for publication bias. Additionally, they did not account for model uncertainty regarding the inclusion of random effects in their meta-analysis. As an alternative, we planned to use the R package `RoBMA` (Bartoš et al. 2023) to estimate an ensemble of meta-analytic models and then use Bayesian model averaging to combine the effect-size estimates of all models based on their posterior probability.[3] Unfortunately, we were unable to estimate these models. We therefore omit this part of the robustness reproduction.

[3] We erroneously specified that we will use the "best-fitting" models for publication bias analyses in our pre-analysis plan, but these models are not the focus of the publication bias analyses.

3.2 Use different model weights for model comparison.

As stated above, it was unclear whether the authors used loo or stacking weights to compare different meta-regression models. In addition to computing the loo weights as stated above, we intended to use posterior model probabilities obtained with bridge sampling in the `brms` package (Bürkner 2017), stacking weights, and the WAIC information criterion (also obtained with `brms`) to compare models. We compared both the selection of the best model and the ordering of models to the corresponding results of the original paper. We defined the same model order as a full reproduction and the same best model (while others may differ in their rank) as a partial reproduction. However, after conducting the analyses, we decided that the definition of a full reproduction is very strict. For example, a change in the order of the third and fourth best models due to some minor numerical difference is hardly relevant to the overall results.[4]

Table 3 presents a summary of the results. For most model weights and outcomes, a parsimonious model version containing only a single predictor was chosen as the best-performing model. When using WAIC weights for f_0 and articulation rate, WAIC weights led to a different best model compared to the other model weight assessments. For the other outcomes, all model weight assessments agreed on the best model. When calculating stacking weights, some Pareto k diagnostic values were too high, indicating that importance sampling was unreliable. We ignored these warnings as there was no information in the original code or manuscript on how to deal with these problems.

3.3 Using different sampling settings for the most complex models

The authors used relatively unconventional settings for the Hamiltonian Monte Carlo (HMC) No U-Turn sampler (NUTS) in Stan (Stan Development Team 2023), as they set `adapt_delta` $= 0.999$ and the `max_treedepth` $= 20$. Choosing a higher

[4] Counter to our initial plan, we did not compute posterior model probability because of computational effort, and because we deemed our current analyses as informative enough.

adapt_delta can decrease divergent transitions and increase computational cost, while the need to increase the max_treedepth could be a sign of model misspecification (see the Stan user guide by the Stan Development Team (2023) for more information). We changed these sampling settings to the brms default (adapt_delta = 0.8, max_treedepth = 10) for the models presented in Table 1 of their manuscript. We then inspected model convergence summaries (such as $\hat{R}$) to check for possible problems in the model specification (e.g., lack of convergence or identifiability of certain parameters). We also preregistered that we would further investigate models with convergence issues. As in the original paper, all of the following models were estimated on 20 imputed data sets. Posterior samples were then aggregated into one final model object.

3.3.1 f_0 Re-estimation of the model with language, age, task, and environment as predictors led to 1,589 divergent transitions after warmup, which indicates severe convergence issues.

3.3.2 VSA Re-estimation of the model with age and language as predictors led to no divergent transitions after warmup. Therefore, we did not investigate this model further.

3.3.3 AR Re-estimation of the model with task, age, and language as predictors led to 32 divergent transitions after warmup.

3.3.4 VD Re-estimation of the model with age and language as predictors led to 114 divergent transitions after warmup.

3.3.5 f_0 Variability Re-estimation of the model with task, age, and language as predictors led to 1,509 transitions after warmup.

It was beyond the scope of this robustness analysis to try to exactly pinpoint the reasons for divergent transitions and other sampling problems for each of the estimated models. Overall, the estimation of a three-level hierarchical structure and a fixed effect for language with a large number of factor levels seem to have been at the upper limit of possible complexity given the available data.

For one example outcome (f_0), we performed a robustness analysis where we attempted to find potential problems in the originally specified model, resolve them, and estimate the revised model with less informative priors and default sampler settings. A full summary of this analysis can be found in the online supplement. Briefly, we iteratively simplified the random effects structure, which enabled us to estimate the model without divergent transitions or obvious convergence problems. The direction and rough magnitude of the effects stayed the same compared to the original model, which gives us additional confidence in the robustness of the effects. Only the evidence for each effect (as operationalized with the Bayes factor) was weaker than in the original model.

4 Discussion

We performed a robustness reproduction of Cox et al. (2023) and could reproduce the main findings of their paper (i.e., estimated effect sizes). However, we were not successful in reproducing the selection of "robust predictors" and the model weight assessment. The analyses by Cox et al. (2023) show several methodological strengths. In general, the code and supplementary materials were more extensive and better organized than in the majority of papers we have encountered in the quantitative social sciences. The Bayesian analyses were performed with care and included a variety of useful and convincing sensitivity analyses (for example, for prior sensitivity).

While the extensive documentation in an RMarkdown file was helpful, the total code, with more than 4,000 lines of code in a single file, was at times difficult for an outsider to understand. In future projects, it may be worth using functional programming to avoid large chunks of similar, copied model code (e.g., for manually selecting different subsets of predictors). Some degree of automation of the analyses over a grid of outcomes and predictor combinations might also have avoided some inconsistencies or minor errors, such as missing code for some model versions, unclear assessment of model weights, or an unexplained change in the number of MCMC iterations. At the same time, the inclusion of data-cleaning steps in the code repository would be helpful and enhance reproducibility.

Concerning reporting standards of statistical analyses, clarification of ambiguous terminology such as "robust predictors" and corresponding references in the analysis scripts (e.g., which analysis produces which table) would enable the reproducibility of all major analyses.

Overall, as our robustness analyses showed, the meta-analysis models had a rather complex structure that was prone to some computational issues. However, even when simplifying the model, the overall results of the original paper remained robust. We congratulate Cox et al. (2023) on their impressive efforts, extensive supplementary material, and reproducible main results. We hope that our suggestions for improved reproducibility and our robustness analyses strengthen the understanding of the current work and provide inspiration for similar projects in the future.

References

Bartoš, F., Maier, M., Wagenmakers, E.-J., Doucouliagos, H. and Stanley, T. D.: 2023, Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods, *Research Synthesis Methods* **14**(1), 99–116.

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

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G. and Fusaroli, R.: 2023, A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech, *Nature Human Behaviour* **7**(1), 114–133.

Mathur, M. B. and VanderWeele, T. J.: 2020, Sensitivity analysis for publication bias in meta-analyses, *Journal of the Royal Statistical Society. Series C, Applied Statistics* **69**(5), 1091–1119.

R Core Team: 2024, R: A Language and Environment for Statistical Computing.
URL: *https://www.R-project.org/*

Stan Development Team: 2023, Stan Modeling Language Users Guide and Reference Manual.
URL: *https://mc-stan.org*

Stan Development Team: 2024, RStan: the R interface to Stan. R package version 2.32.6.
URL: *https://mc-stan.org/*

Ushey, K. and Wickham, H.: 2024, *renv: Project Environments*. R package version 1.0.7.
URL: *https://CRAN.R-project.org/package=renv*

Yao, Y., Vehtari, A., Simpson, D. and Gelman, A.: 2018, Using stacking to average bayesian predictive distributions (with discussion), *Bayesian Analysis* **13**(3).

5 Tables

Table 1: Reproduction Package Contents and Reproducibility

Reproduction Package Item	**Fully**	**Partial**	**No**	**Not applicable**
Raw data provided		✓		
Analysis data provided		✓		
Cleaning code provided			✓	
Analysis code provided		✓		
Reproducible from raw data				✓
Reproducible from analysis data		✓		

Notes: This table summarizes the reproduction package contents contained in Cox et al. (2023).

Table 2: Comparison of original and reproducibility attempt

Feature	**Original Study**	**Reproducibility Study**
Average Effect Size		
f_0	1.19 (0.81, 1.58)	1.20 (0.81, 1.59)
f_0 variability	0.46 (0.21, 0.71)	0.49 (0.20, 0.77)
Vowel space area	0.81 (0.44, 1.16)	0.80 (0.43, 1.18)
Articulation rate	-1.11 (-1.80, -0.39)	-1.20 (-1.89, -0.48)
Vowel duration	0.51 (0.16, 0.86)	0.50 (0.16, 0.86)
Evidence Ratio		
f_0	Inf	Inf
f_0 variability	817.18	408.09
Vowel space area	1,799	1,499
Articulation rate	390.3	271.73
Vowel duration	67.7	77.95
Study s.d.		
f_0	0.91 (0.72, 1.14)	0.92 (0.73, 1.15)
f_0 variability	0.76 (0.60, 0.95)	0.76 (0.59, 0.95)
Vowel space area	0.61 (0.41, 0.86)	0.61 (0.41, 0.86)
Articulation rate	0.74 (0.42, 1.19)	0.75 (0.43, 1.19)
Vowel duration	0.50 (0.12, 0.92)	0.51 (0.14, 0.92)
Robust predictors		
f_0	Language, age, task, environment	?
f_0 variability	Language, task	?
Vowel space area	Language	?
Articulation rate	Language, age, task	?
Vowel duration	Language, age	?

Notes: As in Cox et al. (2023, p.118), the average effect size refers to "the average effect size across infant ages and languages in the best model for the acoustic measure". Numbers in parentheses indicate the 95% Credible Interval.

Table 3: Comparison of different model weight assessments

Outcome	**Weights**	**Intercept**	**Environment**	**Task**	**Age**	**Language**	**AL**	**ELA**	**TLA**	**TELA**
F0	loo	0.019 (3)	0.008 (4)	**0.808 (1)**	0.157 (2)	0.000 (9)	0.000 (7)	0.000 (8)	0.007 (5)	0.001 (6)
F0	stacking	0.000 (7)	0.000 (4)	**0.555 (1)**	0.409 (2)	0.000 (5)	0.000 (9)	0.000 (6)	0.000 (8)	0.036 (3)
F0	WAIC	0.005 (8)	0.026 (7)	0.064 (5)	0.141 (3)	0.001 (9)	0.040 (6)	0.103 (4)	0.214 (2)	**0.405 (1)**
VSA	loo	**0.393 (1)**	0.222 (2)	0.173 (3)	0.125 (4)	0.073 (5)	0.004 (7)	0.006 (6)	0.003 (9)	0.003 (8)
VSA	stacking	**0.768 (1)**	0.000 (4)	0.000 (3)	0.000 (5)	0.232 (2)	0.000 (6)	0.000 (7)	0.000 (9)	0.000 (8)
VSA	WAIC	**0.229 (1)**	0.170 (3)	0.146 (4)	0.078 (5)	0.218 (2)	0.046 (7)	0.048 (6)	0.030 (9)	0.035 (8)
AR	loo	0.000 (6)	0.000 (3)	0.000 (7)	0.001 (2)	0.000 (8)	0.000 (4)	0.000 (9)	**0.999 (1)**	0.000 (5)
AR	WAIC	0.025 (3)	0.024 (4)	0.010 (9)	0.028 (2)	0.014 (8)	0.024 (5)	0.021 (7)	**0.833 (1)**	0.022 (6)
AR	stacking	0.000 (8)	0.286 (2)	0.000 (4)	**0.714 (1)**	0.000 (7)	0.000 (9)	0.000 (6)	0.000 (3)	0.000 (5)
F0V	loo	0.221 (2)	0.082 (3)	**0.657 (1)**	0.039 (4)	0.000 (7)	0.000 (8)	0.000 (9)	0.000 (5)	0.000 (6)
F0V	stacking	0.334 (2)	0.000 (3)	**0.665 (1)**	0.000 (4)	0.000 (6)	0.000 (7)	0.000 (9)	0.000 (8)	0.000 (5)
F0V	WAIC	0.046 (2)	0.032 (3)	**0.869 (1)**	0.019 (4)	0.003 (7)	0.001 (8)	0.001 (9)	0.016 (5)	0.012 (6)
VD	loo	0.045 (7)	0.044 (8)	0.032 (9)	0.072 (5)	**0.318 (1)**	0.206 (2)	0.089 (4)	0.136 (3)	0.056 (6)
VD	stacking	0.000 (5)	0.000 (6)	0.000 (9)	0.000 (3)	**0.919 (1)**	0.081 (2)	0.000 (8)	0.000 (7)	0.000 (4)
VD	WAIC	0.011 (9)	0.014 (7)	0.011 (8)	0.021 (6)	**0.344 (1)**	0.169 (3)	0.133 (4)	0.170 (2)	0.127 (5)

Notes: Model weights for models with different predictors, with the model-specific rank for each outcome-weight combination in parentheses. Abbreviated model titles represent models with multiple predictors based on their first letter. For example, “TELA” represents the model with task, age, language, and environment as predictors. The best model for each outcome-weight combination is presented in bold.