OPEN MIND

Discoveries in Cognitive Science

an open access journal

Citation: Zettersten, M., Cox, C., Bergmann, C., Tsui, A. S. M., Soderstrom, M., Mayor, J., Lundwall, R. A., Lewis, M., Kosie, J. E., Kartushina, N., Fusaroli, R., Frank, M. C., Byers-Heinlein, K., Black, A. K., & Mathur, M. B. (2024). Evidence for Infant-directed Speech Preference Is Consistent Across Large-scale, Multi-site Replication and Meta-analysis. *Open Mind: Discoveries in Cognitive Science*, *8*, 439–461. https://doi.org/10.1162/opmi_a_00134

DOI:
https://doi.org/10.1162/opmi_a_00134

Supplemental Materials:
https://doi.org/10.1162/opmi_a_00134

Received: 17 October 2023
Accepted: 19 February 2024

Competing Interests: The authors declare no conflict of interests.

Corresponding Author:
Martin Zettersten
martincz@princeton.edu

The MIT Press

REPORT

Evidence for Infant-directed Speech Preference Is Consistent Across Large-scale, Multi-site Replication and Meta-analysis

Martin Zettersten[1]*, Christopher Cox[2]*, Christina Bergmann[3]*, Angeline Sin Mei Tsui[4], Melanie Soderstrom[5], Julien Mayor[6], Rebecca A. Lundwall[7], Molly Lewis[8], Jessica E. Kosie[1], Natalia Kartushina[6], Riccardo Fusaroli[2], Michael C. Frank[4], Krista Byers-Heinlein[9], Alexis K. Black[10], and Maya B. Mathur[11]

[1]Department of Psychology, Princeton University
[2]Department of Linguistics, Cognitive Science and Semiotics, School of Communication and Culture, Aarhus University; Interacting Minds Center, School of Culture and Society, Aarhus University
[3]Osnabrück University of Applied Sciences
[4]Department of Psychology, Stanford University
[5]Department of Psychology, University of Manitoba
[6]Department of Linguistics and Scandinavian Studies, University of Oslo
[7]Psychology Department and Neuroscience Center, Brigham Young University
[8]Department of Psychology/Social and Decision Sciences, Carnegie Mellon University
[9]Department of Psychology, Concordia University
[10]School of Audiology and Speech Sciences, University of British Columbia
[11]Quantitative Sciences Unit, Stanford University
*Share joint first-authorship.

Keywords: infant-directed speech, meta-analysis, mega-analysis, multi-lab replication, looking time preference

ABSTRACT

There is substantial evidence that infants prefer infant-directed speech (IDS) to adult-directed speech (ADS). The strongest evidence for this claim has come from two large-scale investigations: i) a community-augmented meta-analysis of published behavioral studies and ii) a large-scale multi-lab replication study. In this paper, we aim to improve our understanding of the IDS preference and its boundary conditions by combining and comparing these two data sources across key population and design characteristics of the underlying studies. Our analyses reveal that both the meta-analysis and multi-lab replication show moderate effect sizes ($d \approx 0.35$ for each estimate) and that both of these effects persist when relevant study-level moderators are added to the models (i.e., experimental methods, infant ages, and native languages). However, while the overall effect size estimates were similar, the two sources diverged in the effects of key moderators: both infant age and experimental method predicted IDS preference in the multi-lab replication study, but showed no effect in the meta-analysis. These results demonstrate that the IDS preference generalizes across a variety of experimental conditions and sampling characteristics, while simultaneously identifying key differences in the empirical picture offered by each source individually and pinpointing areas where substantial uncertainty remains about the influence of theoretically central moderators on IDS preference. Overall, our results show how meta-analyses and multi-lab replications can be used in tandem to understand the robustness and generalizability of developmental phenomena.

INTRODUCTION

Across many cultures, adults adjust the way they speak with infants compared to how they speak with other adults (Cox et al., 2023; Fernald et al., 1989; Hilton et al., 2022). This type of speech addressed to infants (infant-directed speech, or IDS) has unique acoustic and linguistic characteristics compared with adult-directed speech (ADS): for example, IDS tends to be produced with a slower articulation rate, a greater degree of pitch variability, and acoustically exaggerated vowels (Hilton et al., 2022; Kalashnikova & Burnham, 2018; Singh et al., 2002; Stern et al., 1983). Decades of research have investigated infants' responsiveness to this distinctive style of speech, finding that infants prefer IDS over ADS from a young age (Cooper & Aslin, 1990; Fernald & Kuhl, 1987; Pegg et al., 1992; Werker & McLeod, 1989) and that this preference persists even when the speech is filtered to contain only prosodic information (Fernald & Kuhl, 1987) or when presented in a foreign language (The ManyBabies Consortium, 2020). IDS has been argued to play an important role in supporting early language and cognitive development, with the speech style initially serving primarily to draw infants' attention, modulate their temperament and express affect, and later serving more specific linguistic and non-linguistic purposes (Cox et al., 2023; Csibra & Gergely, 2009; Eaves et al., 2016; Fernald et al., 1989; Hartman et al., 2017; Peter et al., 2016; Snow & Ferguson, 1977; Soderstrom, 2007).

Given its centrality in theories of language and cognitive development, how robust is the evidence for infants' IDS preference? A substantial body of research on the IDS preference has culminated in both i) a community-augmented meta-analysis (MA) of published behavioral studies (Anderson et al., 2021; Dunst et al., 2012) and ii) a multi-lab replication (MLR) study (The ManyBabies Consortium, 2020). How can we compare and synthesize the findings from these two different types of data sources? The aim of this paper was to improve our understanding of the relationship between MA and MLR evidence and to determine the generalizability and boundary conditions of the IDS effect across theoretically relevant dimensions.

The first source of evidence we considered was a community-augmented MA of the IDS preference—a meta-analysis with openly accessible data that can be dynamically updated through new contributions from the research community (Tsuji et al., 2014). This MA was developed based on a previously published MA that analyzed 16 papers with a total of 51 effect sizes published between 1983 and 2011 (Dunst et al., 2012). In the published meta-analysis, infants generally preferred to listen to IDS over ADS speech stimuli (Cohen's d = 0.67, 95% CI [0.57, 0.76]). This report also documented variability across several moderators, including that (i) older infants exhibited a stronger preference to attend to IDS over ADS than younger infants and (ii) that characteristics of the methodological design and stimuli systematically affected IDS preference (e.g., effects were stronger if speakers were unfamiliar to infants). The original Dunst et al. MA was subsequently revised and augmented (see Methods for details) by the MetaLab community of infant researchers, resulting in a MA encompassing 30 papers published between 1985 and 2020 that contributed a total of 112 effect sizes (https://metalab.stanford.edu; Anderson et al., 2021; Bergmann et al., 2018). Notably, this community-augmented meta-analysis resulted in a substantially smaller effect size estimate (d = 0.35, 95% CI [0.22, 0.47]).

Our second source of evidence was a MLR of IDS preference, in which 69 laboratories on four continents (Asia, Australia, Europe, North America) collected data from over 2700 infants aged between 3 and 15 months (The ManyBabies Consortium, 2020). The general aim of ManyBabies 1 was to replicate the main phenomenon of IDS preference among infants while

assessing the impact of several theoretically meaningful variables, including infant age, language experience and testing methods. IDS preference was measured by analyzing infants' behavioural visual responses to IDS and ADS speech stimuli using three different methods that were self-selected by each participating lab: central fixation, the head-turn preference procedure (HPP) and eye-tracking. The sets of ADS and IDS stimuli were held constant across laboratories and were created by recording a small number of North American mothers in a semi-naturalistic speech elicitation task. The results from the MLR indicated i) that infants generally prefer to listen to IDS over ADS speech stimuli (overall Cohen's d = 0.35, 95% CI [0.29, 0.42]), ii) that the preference for IDS over ADS was strongest in the oldest age range tested, iii) that infants learning North American English (i.e., whose native language matched that of the test stimuli) showed stronger effects than those learning languages other than North American English, and iv) that the HPP elicited stronger effects than both central fixation and eye-tracking.

On the surface, the evidence for the IDS preference appears broadly consistent across the MLR and the community-augmented MA: both sources show small-to-moderate positive effect sizes. However, these studies take fundamentally different approaches to deriving their overall estimates, with distinctive strengths and weaknesses. MAs have traditionally been considered a gold standard form of evidence, by offering a bird's-eye view of the generalizability of a phenomenon—as well as the heterogeneity of effects—across a variety of designs and populations. However, they have also been criticized on several grounds (Corker, 2022; Lakens et al., 2016; Siddaway et al., 2019; Stanley, 2001). One major concern is that MAs are subject to publication bias (Sterne et al., 2001). MAs are often limited to the available (published and grey) literature, and a small set of unpublished studies individual researchers are willing and able to dredge from the file drawer. This limitation may bias estimates, as positive results are typically over-represented in the published literature (Masicampo & Lalande, 2012; Mathur & VanderWeele, 2021a; McShane & Gal, 2017; Sterne et al., 2001). A second concern is that heterogeneity in the studies included in a meta-analysis can threaten to complicate practical interpretation when taken to an extreme, i.e., meta-analyses may be comparing "apples to oranges" (Eysenck, 1978; Simonsohn et al., 2022). While there are statistical approaches that attempt to correct for publication bias and measure and account for heterogeneity, major concerns about the validity of MA results—even when corrected—remain.

In part due to the limitations of MAs, many researchers have begun to consider MLRs the new gold standard. In such designs, multiple labs conduct replications of original studies by implementing a common experimental protocol to test a research question across sites (e.g., Ebersole et al., 2016, 2020; Jones et al., 2021; Klein et al., 2014, 2018, 2019; The ManyBabies Consortium, 2020). Like MAs, MLRs (such as ManyBabies 1) can achieve larger aggregated sample sizes than are typical in single-lab studies, but the similarity in implementation across labs may offer greater comparability within the dataset. Moreover, MLRs do not suffer from concerns about publication bias, because the results from all labs are reported transparently regardless of outcome. On the other hand, more uniformity in experiment implementation may lead to effect size estimates that are less robust to methodological and analytical differences; that is, the measured effect size may reflect the particular methodological and analytic choices of the study. Therefore, more narrowly defined experimental parameters may limit the degree to which MLRs can speak to the generalizability and boundary conditions of a phenomenon (Visser et al., 2022; Yarkoni, 2020).

While both MAs and MLRs individually represent valuable methods for estimating an effect of interest, consulting either a MA or MLR in isolation likely provides an incomplete picture of

theoretically important phenomena. Furthermore, past work comparing MAs and MLRs suggests that the results obtained from these two approaches often do not agree. In a study that systematically compared 15 pairs of published MAs and MLRs within the field of psychology, Kvarven et al. (2020) found significant differences in mean effect sizes for 12 of the pairs, with MA effect sizes on average three times the size of those obtained via MLRs. What drives these differences remains unclear. For example, in a reanalysis of the same data, Lewis et al. (2022) concluded that these discrepancies could not be fully accounted for by publication bias. An alternative explanation appeals to potential heterogeneity in the MA (Lewis et al., 2022). If there is true heterogeneity in the studies summarized in the MA, this could create a false impression of inconsistent results between the MA and the MLR, despite the MLR estimate falling within a reasonable subdistribution of effects in the MA. Given the limitations of MAs and MLRs considered alone—and resulting divergences in the conclusions derived from each method—, a promising approach to understanding a key phenomenon of interest is to combine and synthesize evidence from both sources. This strategy seems particularly useful given how the benefits of each approach may counteract the limitations of the other. MLRs can provide estimates that do not suffer from publication bias, whereas MAs can typically offer estimates across a wider variety of experimental design choices than MLRs.

In the current paper, we investigate the overall magnitude, generalizability, and boundary conditions of the IDS preference effect by integrating and comparing experiment-level data from both the MA and MLR. Unlike past comparisons focusing on the overall effect size of MAs and MLRs (e.g., Kvarven et al., 2020), we explicitly model data from the individual studies included in the MA and individual experiments contributing to the MLR. We simultaneously code key features of each experiment to investigate whether heterogeneity in effects across moderating variables thought to substantially impact IDS preference can explain any discrepancies between the MA and MLR. We take a meta-regression approach, estimating the magnitude of IDS preference aggregating across the two data sources with and without theoretically-motivated moderator-level variables. Together, these analyses increase our overall understanding of IDS preference while also providing a detailed case study of the relationship between MA and MLR. We focus on three main questions:

1. Do the MA and MLR provide comparable estimates of infant preference for IDS?
2. Does accounting for study-level moderators and publication bias affect the comparison of the estimates across the two approaches?
3. Are there differences between the MA and the MLR in how study-level moderators predict IDS preference?

The first two questions followed a preregistered analytic approach, while the third question was investigated in additional exploratory analyses. The preregistered analyses were designed to be conducted using the original Dunst et al. (2012) meta-analysis as the main MA source. However, after the preregistered plan was finalized, two key events occurred: (a) we uncovered substantial issues with coding decisions in the original meta-analysis that required revision and (b) the original meta-analysis was augmented via systematic search to include almost twice the number of studies (see Methods and Supplementary Materials). In order to test our primary research questions with the most extensive and accurate evidence source possible, we therefore opted to deviate from the preregistration and execute our preregistered analytic plan using the community-augmented MA as our primary meta-analytic data source.[1]

[1] For parallel analyses using both the original meta-analysis and a revised version of the meta-analysis, as well as a discussion of discrepancies, see Supplementary Materials (Section 5 and Section 6).

METHODS

All confirmatory analyses were preregistered prior to data analysis at https://osf.io/scg9z. The Supplementary Materials provide further details on the preregistration framework (Section 1.1) and deviations from our preregistered plan (Section 1.2), and contextualizes the updates to datasets (Section 6).

Meta-analysis

The Original Dunst et al. (2012) Meta-analysis. The MA by Dunst et al. (2012) reports study-level effect sizes in Appendix C of the original study and moderator variables in Appendices A and B. We digitized these variables, and an independent team checked and corrected the resulting spreadsheet to fully reflect the published meta-analysis. We additionally computed effect size variances using standard formulae based on reported standardized mean difference (SMD) and sample sizes. To supplement the MA with moderators that were relevant for the research questions in this study but not reported on in the MA (Dunst et al., 2012), it was necessary to re-examine the papers reporting on the original experiments. This process led to a number of additional moderating variables that included further detail about (1) whether the test language was native for infant participants, non-native, or an artificial language; (2) whether the main question of the study was focused on IDS preference; (3) variation in experimental methods (e.g., whether test trials were infant-controlled or had a fixed duration); and (4) variation in participant exclusions and exclusion criteria (e.g., what number of test trials were required for infant inclusion).

Revisions to the Dunst et al. Meta-analysis. When coding for additional moderators for the studies included in the original MA (Dunst et al., 2012), we encountered substantial issues, such as incorrectly reported effect sizes and inappropriate inclusion and exclusion of experimental conditions (as discussed further in Section 2.1 of the Supplementary Materials). The original MA never underwent a formal peer review process, which could have caught some of these errors; however, even published and reviewed MAs are not exempt from replicability and reproducibility issues (Maassen et al., 2020; Nuijten et al., 2016).

The Community-augmented Meta-analysis. The revised Dunst et al. meta-analysis was subsequently augmented based on new literature searches conducted in 2017 and 2019, resulting in an updated, community-augmented database of studies on infants' IDS preference in MetaLab (https://metalab.stanford.edu/; Tsuji et al., 2014). Further details about the augmentation process are provided in Section 2.2 of the Supplementary Materials. The community-augmented MA comprised k = 30 studies contributing a total of m = 112 estimates, which included a median of n = 16.50 participants.

To provide the most comprehensive, up-to-date point of comparison between the MLR and the MA, we focus our preregistered analyses on the updated, community-augmented MA that includes revisions to the issues identified in the original meta-analysis. All analyses using the original dataset (i.e., Dunst et al., 2012) and a revised version containing only studies included in the original MA (i.e., correcting errors or other issues in the coding of papers from the original MA, but not updating the dataset to include additional studies)—as well as a discussion of differences with the main conclusions presented here—can be found in the Supplementary Materials (Section 5 and Section 6).

Multi-Lab Replication: ManyBabies 1

A total of k = 62 labs contributed a total of m = 102 estimates to the dataset, because single labs could contribute data in multiple age groups. This dataset is identical to the data in the

original analyses (The ManyBabies Consortium, 2020), which excluded infants who did not provide at least one trial per condition (IDS and ADS in paired trials) and labs providing estimates from less than ten infants. Note that slightly fewer labs were included in this analysis in The ManyBabies Consortium (2020) compared to the overall number of labs contributing to the project ($N = 67$) because of stricter inclusion criteria (infants were required to contribute paired IDS and ADS trials). The data were downloaded from the public GitHub repository (https://github.com/manybabies/mb1-analysis-public) of the MLR. Effect sizes were computed, both here and in the original paper, as standardized mean differences (SMD) based on the average looking time difference in IDS and ADS trials divided by the pooled standard deviation of looking time on the level of study (i.e., an age group within a lab); variance was computed accordingly. Post hoc, we added all moderators that were not part of the original dataset, such as speaker identity (e.g., unfamiliar female), to align this dataset with the MA (see Table 1). The estimates in this dataset are based on a median of $n = 16$ participants per age group (ranging from 10 to 46). For further details on the MLR, including participant sampling and exclusion criteria, see sections 3.1 and 3.2 of the Supplementary Materials.

Hypothesized Estimate-level Moderators

In our primary analyses, we investigated eight hypothesized estimate-level moderators of the IDS preference effect, which we coded in both sources (i.e., the MA and the MLR datasets; for an overview, see Table 1; for details, see Section 4.1 of the Supplementary Materials). These comprised one characteristic of the study population (average participant age [in months, mean-centered]), four characteristics of the stimuli (test language, speech type, speaker familiarity, and mode of presentation), two methodological characteristics (experimental method and dependent measure), and an overall estimate characteristic (study goal, i.e., whether infants' preference for infant- over adult-directed speech was the main research question of a paper). One additional moderator we considered was infants' native language. However, infants' native language was heavily skewed towards North American English and is confounded with whether stimuli were presented in infants' own native language, as any non-native stimuli were North American English across both the MA and the MLR. We thus use this factor only for exploratory analyses but mention it here for completeness (cf. also Figure 1 in Section 4.2 of the Supplementary Materials for more information on the distribution of interactions between moderators). In our regression models, we dummy-coded the binary and categorical moderators such that the reference level represented the most common level in the meta-analysis. Similarly, we centered the single continuous moderator, mean age in months, by its mean in the MA.

Statistical Analyses

Evidence Measures. We used three metrics to characterize evidence strength for IDS preference in each source and to compare evidence between the sources. First, we estimated the average effect size (SMD) in each source. Examining the difference between sources in these average effect sizes is an important first step, but this approach can exaggerate differences between meta-analyses if effects are highly heterogeneous. In such cases, a fairly large difference between means can occur simply as a result of heterogeneity. By the same token, heterogeneity might lead two meta-analytic estimates to appear similar despite important differences in the underlying evidence base (Mathur & VanderWeele, 2019). For this reason, we also estimated other metrics of agreement that more holistically compare the distributions of effects rather than only their means. As a second metric of evidence strength, we

Table 1. The distribution of moderators in the community-augmented meta-analysis (MA) and multi-lab replication (MLR).

		MA	MLR
Number of effect sizes		112	102
Infant age (months; centered)	Mean (SD)	0.00 (5.76)	1.22 (3.03)
Test language			
	Native	103 (92.0%)	46 (45.1%)
	Non-Native	6 (5.4%)	56 (54.9%)
	Artificial	3 (2.7%)	0 (0%)
Native language			
	Cantonese	4 (3.6%)	0 (0%)
	Dutch	0 (0%)	5 (4.9%)
	English	103 (92.0%)	62 (60.8%)
	French	0 (0%)	6 (5.9%)
	German	0 (0%)	14 (13.7%)
	Hungarian	0 (0%)	2 (2.0%)
	Italian	0 (0%)	1 (1.0%)
	Japanese	5 (4.5%)	4 (3.9%)
	Korean	0 (0%)	3 (2.9%)
	Norwegian	0 (0%)	1 (1.0%)
	Spanish	0 (0%)	2 (2.0%)
	Swiss German	0 (0%)	1 (1.0%)
	Turkish	0 (0%)	1 (1.0%)
Experimental method			
	Central Fixation	67 (59.8%)	68 (66.7%)
	HPP	39 (34.8%)	34 (33.3%)
	Other	6 (5.4%)	0 (0%)
Speech type			
	Simulated	75 (67.0%)	0 (0%)
	Naturalistic	30 (26.8%)	102 (100%)
	Filtered or Synthesized	7 (6.3%)	0 (0%)
Speaker familiarity (own mother)			
	No	109 (97.3%)	102 (100%)
	Yes	3 (2.7%)	0 (0%)

Table 1. (*continued*)

		MA	MLR
Mode of presentation			
	Audio	93 (83.0%)	102 (100%)
	Video	19 (17.0%)	0 (0%)
Dependent measure			
	Preference	104 (92.9%)	102 (100%)
	Affect	8 (7.1%)	0 (0%)
Main question: IDS preference			
	Yes	88 (78.6%)	102 (100%)
	No	24 (21.4%)	0 (0%)

estimated the percentage of population effects[2] in each source that were positive, representing any preference for IDS regardless of magnitude (Mathur & VanderWeele, 2019, 2020a). As a third metric, for a more stringent assessment, we estimated the percentage of population effects in each source representing only effects that were stronger than $SMD > 0.2$ (Mathur & VanderWeele, 2019, 2020a) (in the predicted direction, i.e., showing an IDS preference).

Between-source Discrepancies Before and After Accounting for Hypothesized Moderators. We fit three meta-regression models predicting effect sizes as standardized mean differences (SMD) in R (R Core Team, 2020).[3]: (1) an **unadjusted model** that compared the two sources (MA and MLR) but did not account for other hypothesized moderators, (2) a **moderated model** that additionally included the other hypothesized moderators, and (3) an exploratory **interaction model** that included the two-way interactions between the moderators and the source of the effect sizes.

The first two models estimated the extent to which the MA and MLR results differed when either ignoring estimate-level moderators (the unadjusted model) or when accounting for them (the moderated model). That is, the unadjusted model estimated average effect sizes for each source, the percentage of positive effects, and the percentage of effects stronger than $SMD = 0.2$ when averaging over the distributions of moderators in each source. In contrast, the moderated model estimated these measures for each source when holding constant all

[2] We use the term "population effects" to refer to population parameters, rather than to point estimates with statistical error.

[3] We used the packages boot (Davison & Hinkley, 1997), table1 (Rich, 2021), MatchIt (Ho et al., 2011), xtable (Dahl et al., 2019), Matrix (Bates & Maechler, 2021), ggplot2 (Wickham, 2016), stringr (Wickham, 2019), forcats (Wickham, 2021a), tidyr (Wickham, 2021b), scales (Wickham et al., 2020), readr (Wickham & Hester, 2020), dplyr (Wickham et al., 2021), testthat (Wickham, 2011), fastDummies (Kaplan, 2020), weightr (Coburn & Vevea, 2019), tableone (Yoshida & Bartel, 2020), renv (Ushey & Wickham, 2021), here (Müller, 2020), tibble (Müller & Wickham, 2021), purrr (Wickham & Henry, 2020), report (Makowski et al., 2023), data.table (Dowle & Srinivasan, 2020), corrr (Kuhn et al., 2020), PublicationBias (Braginsky et al., 2023), metafor (Viechtbauer, 2010), tidyverse (Wickham et al., 2019), knitr (Xie, 2014), and robumeta (Fisher et al., 2017).

moderators to their average values (in the case of continuous variables) or their most common values (in the case of categorical variables) in the meta-analysis, which we used as reference levels. Both models included all $m = 214$ estimates from both data sources. We anticipated that the moderated model (and consequently, the interaction model) would not be statistically estimable if some moderators were relatively highly correlated, so we removed moderators one-by-one in ascending order of scientific relevance until the model was estimable. Three moderators emerged as estimable in the moderated model: infant age, test language, and experimental method (see Supplementary Materials Section 4 for further details).

Finally, we fit an exploratory model including the two-way interactions between source (MA vs. MLR) and the same three estimable moderators in the moderated model (i.e., infant age, test language, and method). We fit this interaction model because each of these three predictors were significantly related to IDS preference in the original ManyBabies analysis (The ManyBabies Consortium, 2020), but did not reach significance in the moderated model. The interaction model thus served to further investigate this discrepancy by estimating the degree to which the effect of each predictor depended on the data source. For this analysis, we simplified the test language predictor (Native vs. Other) and the method variable (HPP vs. Other) into centered, binary variables (as opposed to three-level categorical variables) in order to achieve model convergence. Note that the results from the moderated model above remain unchanged if the moderator variables are simplified in this manner.

Publication Bias. For the MA, we assessed the possible contribution of publication bias to the results and to between-source discrepancies in average effect sizes. First, we assessed publication bias in the MA using selection model methods (Vevea & Hedges, 1995), sensitivity analysis methods (Mathur & VanderWeele, 2020b), and the significance funnel plot (Mathur & VanderWeele, 2020b). These methods assume that the publication process favors "statistically significant" (i.e., $p < 0.05$) and positive results over "nonsignificant" or negative results, an assumption that conforms well to empirical evidence on how publication bias operates in practice (Mathur & VanderWeele, 2021a; Masicampo & Lalande, 2012; McShane & Gal, 2017). We used visual diagnostics to assess the plausibility of these assumptions. "Publication bias" in this context could reflect the aggregation of multiple sources of bias, including, for example, investigators' selective reporting of experiments or preparation of papers for submission as well as journals' selective acceptance of papers.

The sensitivity analysis methods do not estimate the actual severity of publication bias, but rather consider how much results might change under varying degrees of hypothetical publication bias. These methods, unlike the selection model, also accommodate the point estimates' non-independence within articles, do not make distributional assumptions, and do not require a large number of studies (Mathur & VanderWeele, 2020b). Using the sensitivity analysis methods, we estimated the meta-analytic mean under hypothetical worst-case publication bias (i.e., if "statistically significant" positive results were infinitely more likely to be published than "nonsignificant" or negative results). This worst-case estimate arises from meta-analyzing only the observed "nonsignificant" or negative studies and excluding the observed "significant" and positive studies. We also estimated the amount of hypothetical publication bias that would be required to shift the estimate in the MA to match the estimate in the MLR (Mathur & VanderWeele, 2020b). A previous study estimated publication bias to favor affirmative results by a factor of 4.7 on average in a small sample of developmental psychology MAs (Mathur & VanderWeele, 2021a). Consistent with this finding, we conducted a post-hoc analysis estimating the meta-analytic mean assuming the same level of publication bias.

RESULTS

Meta-analysis and MLR Results Modeled Separately

The overall effect size in the MA dataset was *SMD* = 0.35 [0.22, 0.47] ($p < 0.0001$), with considerable heterogeneity (estimated standard deviation of population effects $\hat{\tau} = 0.31$). This effect size was roughly half the size of the effect size for IDS preference reported in the original MA (Cohen's $d = 0.67$) by Dunst et al. (2012) (cf., Section 5.1 and 5.2 of the Supplementary Materials), indicating a substantial effect of the revisions and extensions performed as part of the community-augmented meta-analysis process (Tsuji et al., 2014). We estimated that the vast majority of the population effects were positive (86% [83%, 90%]) and that most effects were stronger than *SMD* = 0.2 (64% [61%, 73%]; Table 2). Among only the MLR studies, the estimated average effect size was *SMD* = 0.34 [0.27, 0.42]; $p < 0.0001$) and with less estimated heterogeneity ($\hat{\tau} = 0.11$) compared to the MA (cf., similar results when applying more stringent participant inclusion criteria on the MLR in Section 5.4.6 of the Supplementary Materials). Descriptively, the meta-analytic effect size in the revised MA was therefore virtually identical to that of the MLR when estimating each effect size separately. For the MLR, we estimated that nearly all of the population effects were positive (100% [96%, 100%]) and that a large majority were stronger than *SMD* = 0.2 (89% [76%, 100%]; Table 2). These results are visualised in Figure 1, which shows positive population effects for studies in both sources, but with the MLR exhibiting more concentration around its average effect size estimate than the MA (see also Figure 7 in the Supplementary Materials Section 5.4.3 for a visualization of the estimated densities of population effects, illustrating the greater heterogeneity of the MA as compared to the MLR).

Table 2. $\hat{\mu}$: Average effect size (*SMD*), as estimated in a meta-regression model containing both sources. % effects > 0: Estimated percentage of positive population effects, as estimated in a meta-analysis or meta-regression model containing one source. % effects > 0.2: Estimated percentage of population effects stronger than *SMD* = 0.2. Discrepancies are calculated by subtracting between each statistical measure in the MLR from that in the MA, such that positive discrepancies indicate larger effect sizes in MA. Bracketed values are 95% confidence intervals, which are model-based for the $\hat{\mu}$ measures (Hedges et al., 2010) and for differences in $\hat{\mu}$ between sources and are bootstrapped for the percentage measures and for all cross-model comparisons (Mathur & VanderWeele, 2020a, 2021b). Confidence intervals are omitted when they were not statistically estimable (i.e., for percentage estimates that were very close to 0% or 100%).

Statistical measure	Unadjusted model	Moderated model
$\hat{\mu}$ in MA	0.34 [0.22, 0.46]	0.32 [0.16, 0.47]
$\hat{\mu}$ in MLR	0.34 [0.27, 0.42]	0.35 [0.22, 0.47]
$\hat{\mu}$ discrepancy	−0.01 [−0.14, 0.13]	−0.03 [−0.2, 0.14]
% effects > 0 in MA	86 [83, 90]	88 [83, 92]
% effects > 0 in MLR	100 [96, 100]	100
% effects > 0 discrepancy	−14 [−17, −7]	−12 [−17, −8]
% effects > 0.2 in MA	64 [61, 73]	71 [62, 79]
% effects > 0.2 in MLR	89 [76, 100]	100
% effects > 0.2 discrepancy	−25 [−40, −6]	−29 [−38, −21]

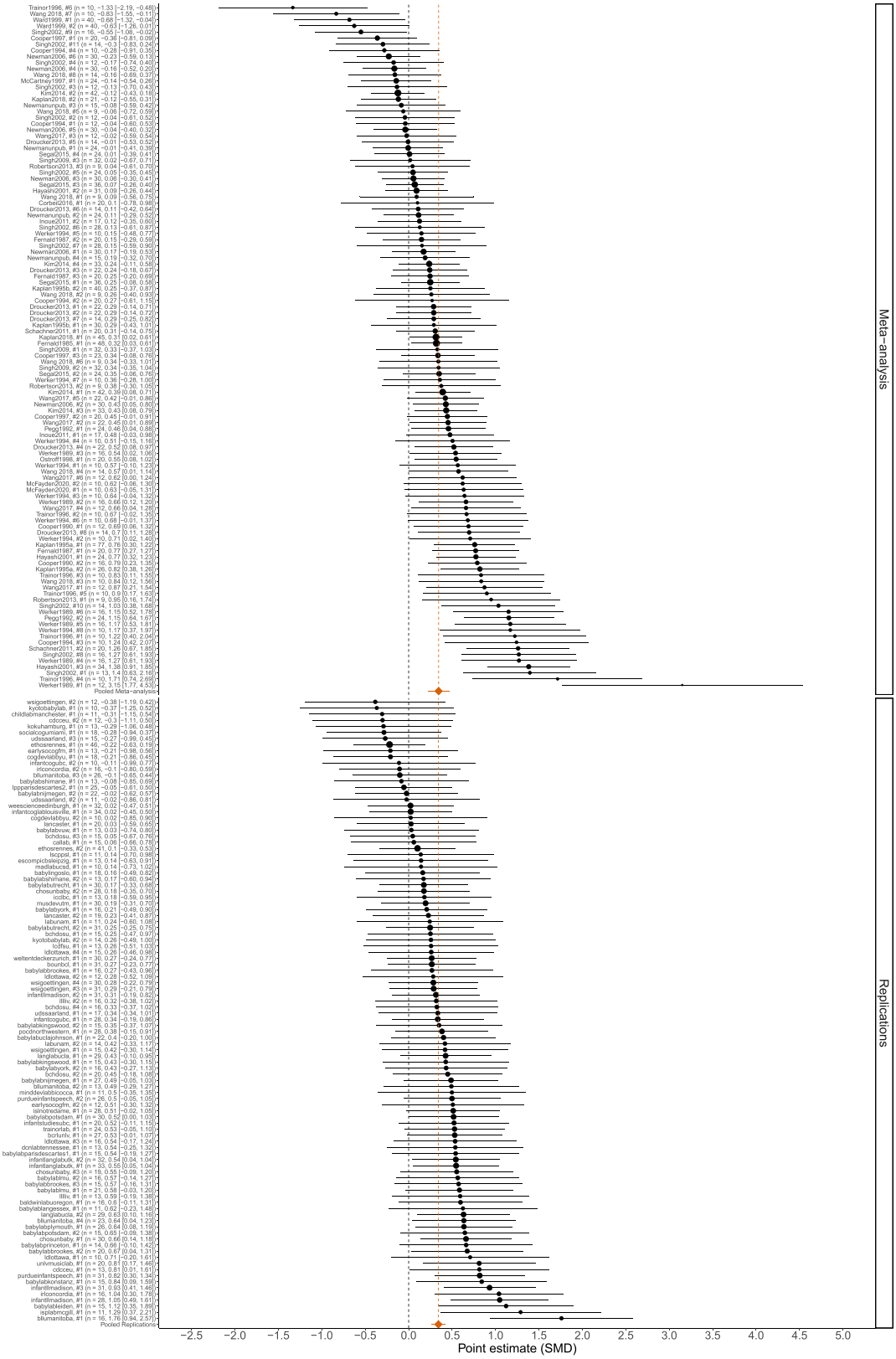

Figure 1. Forest plot of studies' point estimates and 95% confidence intervals in the MA (top panel) and MLR (bottom panel). Orange diamonds: pooled estimates within each source. Dashed vertical line: null.

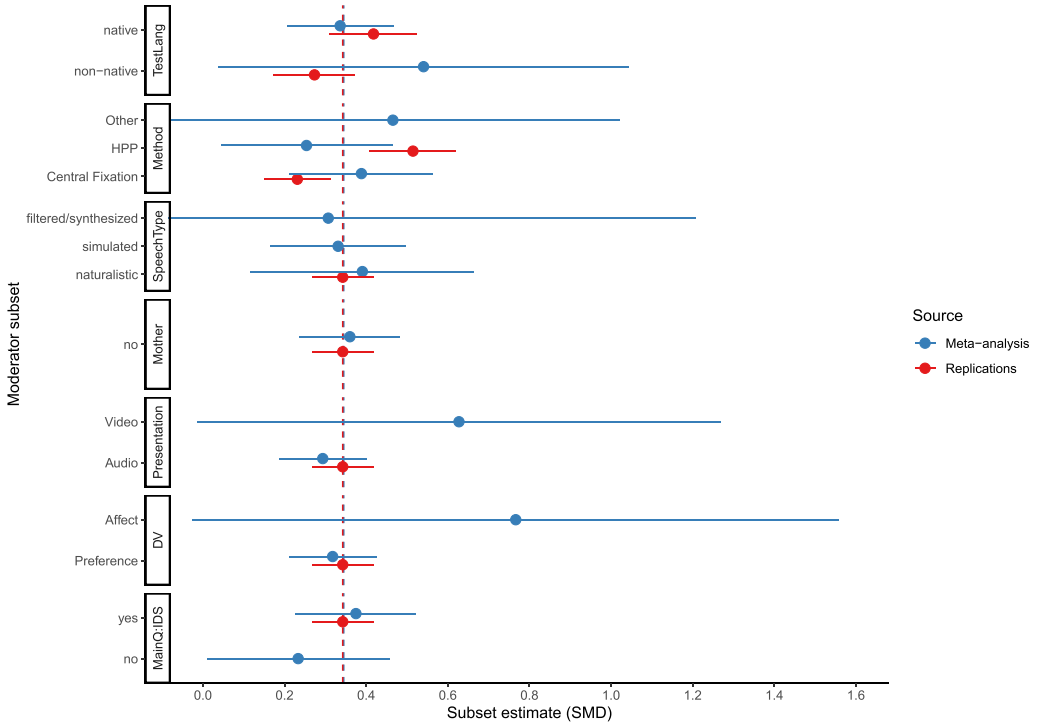

Figure 2. Forest plot showing, for each categorical candidate moderator, the pooled point estimates for the subset of studies in the MA and in the MLR, respectively, with a given level of the moderator (including only levels with at least 5 observations). Error bars are 95% confidence intervals. Error bars for many estimates are wide due to a limited number of observations at certain levels of a given moderator variable. Dashed vertical lines are unadjusted estimates in all MA studies and in all MLR studies. These lines overlap because the two estimates are virtually identical.

To delve further into the moderator analyses, Figure 2 shows, for each categorical candidate moderator, the pooled point estimates for the subset of studies in the MA and in the MLR, respectively, within a given level of the moderator. These simple, post hoc subset analyses stratify on only one moderator at a time and exclude those subsets that could not be estimated (e.g., familiarity of the speaker).

Combined Models

We next considered models combining both the MA and the MLR datasets. We first fit an unadjusted model that combined the two sources without any additional moderators, confirming that effect sizes in the MA did not differ on average from effect sizes in the MLR, −0.01 (95% CI: [−0.14, 0.13]) units on the *SMD* scale (Table 2). There was considerable residual heterogeneity (estimated standard deviation of population effects $\hat{\tau}_{\text{unadjusted}} = 0.27$). Next, we fit a moderated model that explored whether IDS preference varied as a function of a set of theoretically meaningful predictor variables. The moderated model converged when we included three moderators besides source: infant age, test language, and method. Table 3 summarizes the estimates of the meta-regression for those remaining moderators. The estimated average effect size in the MA and in the MLR when setting the moderators to their average value (in the case of the continuous moderator infant age) or their most common value (in the case of the two categorical moderators; method: central fixation, test language: native) in the MA was, respectively, 0.32 [0.16, 0.47] and 0.35 [0.22, 0.47]. Thus, we also did not observe a significant difference between the effect sizes estimated for the MA and the MLR when controlling for moderators of theoretical interest, −0.03 [−0.20, 0.14]. Moreover, none of the three moderator variables showed a significant effect on the magnitude of IDS preference across the MA and the MLR. The residual heterogeneity increased slightly relative to the

Table 3. Meta-regression estimates of moderation by various study design and participant characteristics. Intercept: estimated mean SMD when all listed moderators are set to 0 (for continuous moderators, the average value in the MA or, for categorical moderators, the most common value in the MA). The estimate of the categorical factor Meta-Analysis represents the change in *SMD* when this factor is true vs not. For infant age, the estimate represents the increase in effect size associated with a 1-month increase in mean infant age. For categorical moderators, estimates represent the increase compared to the reference level (Test Language: Native, and Method: Central Fixation, respectively). Bracketed values are 95% confidence intervals. *p*-values represent tests of moderators' coefficients themselves (vs. 0) in the meta-regression.

Moderator	Est	CI	*p*-value
Intercept	0.35	[0.22, 0.47]	< 0.0001
Source: Meta-Analysis	−0.03	[−0.2, 0.14]	0.709
Infant Age (months)	0.01	[−0.00, 0.03]	0.120
Test Language: Non-native	−0.06	[−0.20, 0.09]	0.427
Test Language: Other	−0.17	[−2.68, 2.34]	0.544
Method: HPP	0.04	[−0.13, 0.21]	0.623
Method: Other	0.28	[−1.86, 2.42]	0.402

unadjusted model ($\hat{\tau}_{\text{mod}} = 0.30$). Overall, IDS preference was estimated to be stable across the data source, method, test language, and infant age. However, many of the confidence intervals were wide, indicating substantial uncertainty about moderation strength.

Finally, we conducted an exploratory analysis in which we included the two-way interaction between source (MA vs. MLR) and each of the three moderator variables (infant age, test language, and method). The results from this model are summarized in Table 4. We found evidence for two key interactions. First, there was a significant interaction between source and infant age ($b = -0.04$ [−0.07, −0.02]; $p = 0.002$). This interaction was driven by the fact that there was a robust increase in the magnitude of the IDS effect across infant age in the MLR ($b = 0.04$ [0.02, 0.07]; $p = 0.0004$), but no appreciable change in IDS across infant age in the MA ($b = 0$ [−0.02, 0.02]; $p = 0.82$; Figure 3A). Second, we found a significant interaction

Table 4. Meta-regression estimates of the moderator interaction model. Intercept: estimated mean SMD when averaging across all (centered) moderators. Age (in months) is mean-centered. Test Language (Native vs. Other) and Method (HPP vs. Other) are treated as binary variables and centered. Bracketed values are 95% confidence intervals.

Moderator	Est	CI	*p*-value
Intercept	0.34	[0.25, 0.42]	< 0.0001
Source (centered)	0.03	[−0.14, 0.20]	0.656
Age (months; centered)	0.02	[0.01, 0.04]	0.001
Test Language (Native vs. Other; centered)	0	[−0.17, 0.17]	0.973
Method (HPP vs. Other; centered)	0.05	[−0.08, 0.17]	0.467
Source * Age	−0.04	[−0.07, −0.02]	0.002
Source * Test Language	−0.18	[−0.52, 0.15]	0.214
Source * Method	−0.38	[−0.63, −0.12]	0.005

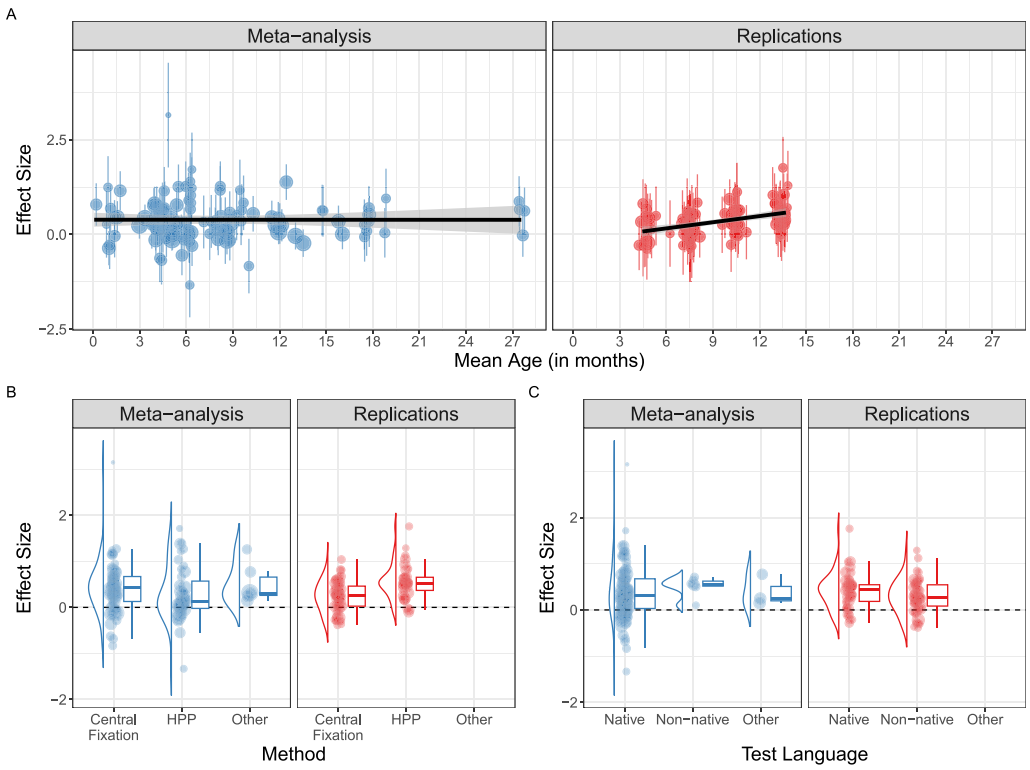

Figure 3. Overview of the distribution of effect sizes in the meta-analysis (MA) and replications (MLR) for three key moderators: infant age (A), method (B), and test language (C). In (A), the black line represents a linear fit through the effect sizes for each source and error bars for individual estimates are 95% confidence intervals.

between source and method (b = −0.38 [−0.63, −0.12]; p = 0.005). Here, the interaction appeared to be driven by a stronger effect of HPP (vs. other methods) in the MLR (b = 0.24 [0.13, 0.34]; $p < 0.0001$), but a numerically opposite, though not significant, effect of method in the MA (b = −0.14 [−0.39, 0.11]; p = 0.24; Figure 3B). There was no interaction between test language and source (b = −0.18 [−0.52, 0.15]; p = 0.21); however, both of these confidence intervals are wide, so moderate to strong moderation effects cannot be ruled out. Residual heterogeneity remained substantial ($\hat{\tau}_{\text{mod}} = 0.27$) but was reduced relative to the combined moderated model. We also assessed the robustness of the results by restricting the MA only to studies with average ages within the range observed in the MLR (3- to 15-month-old infants). We found broadly comparable results for the interaction between source and method and source and age, albeit with increased uncertainty (see Section 5.4.7 of the Supplementary Materials).

Publication Bias

We also considered the extent to which publication bias may be affecting estimates of differences between the MA and MLR. The MA contained 41 affirmative (i.e., statistically significant and positive-signed) and 71 nonaffirmative studies. We began by implementing a correction for publication bias, estimating the selection ratio from the MA itself. Based on the MA, we estimated that affirmative results were favored by a factor of 1.5. The average effect size in the MA after correction was SMD = 0.28 [0.14, 0.41]; $p < 0.0001$ (Vevea & Hedges, 1995), which was indeed somewhat smaller than the uncorrected estimate of SMD = 0.35. Next we applied sensitivity analyses for publication bias, considering what the true effect size would be under several different scenarios. Under hypothetical worst-case publication bias (i.e., if "statistically

significant" positive results were infinitely more likely to be published than "nonsignificant" or negative results), the MA mean would decrease to 0.09 [−0.01, 0.18], which was significantly less than the estimate in the MLR. Under "typical" publication bias in this field (favoring affirmative results by 4.7-fold), the MA average would decrease to 0.17 [0.06, 0.27]. In both cases these estimates were lower than those in the MLR and—in the worst-case scenario—included zero in the 95% CI. Thus the estimate obtained in the MLR is—if anything—likely to be larger than the estimate of the MA under typical or worst-case assumptions about the severity of publication bias (cf., Section 5.4.5 of the Supplementary Materials for additional analyses of publication bias).

DISCUSSION

Infant-directed speech (IDS) and its captivating nature for infants is an important phenomenon for many theories of early linguistic and social development. To improve our understanding of IDS preference and its boundary conditions, we compared and synthesized the evidence from two data sources: a community-augmented meta-analysis (MA) and an extensive multi-lab replication (MLR). Our analyses showed that the overall estimates across the two studies were similar, though the MA exhibited a greater degree of heterogeneity than the MLR. The estimates for the MA and the MLR remained comparable when including a range of theoretically-motivated moderators: adding moderators neither decreased heterogeneity nor produced significant differences between the MLR and MA estimates. However, in exploratory analyses, we found that the predicted effects of key moderators differed between data sources. Specifically, an interaction model showed i) an age-related increase in the strength of the effect in the MLR and no clear developmental change in the MA and ii) a stronger effect for the HPP method (compared to other experimental methods) in the MLR, but not in the MA. Together, these findings show that the MA and MLR provide converging evidence for the IDS preference across a wide range of participant, stimulus, and design characteristics, while also highlighting areas where substantial uncertainty remains about the effect of key moderators on IDS preference.

Implications for Understanding the IDS Preference

Our main finding is that the IDS preference effect generalizes across relevant study dimensions in both the MA and MLR. The moderated models showed convergent results for IDS preference, with infants showing a general preference to attend to IDS over ADS stimuli during early development across a wide variety of ages, task contexts and linguistic backgrounds. This analysis thus conformed to previous studies showing that the unique properties of IDS robustly captivate infants' attention from an early point in development (Cooper & Aslin, 1990; Fernald & Kuhl, 1987; Pegg et al., 1992; Werker & McLeod, 1989). The size of the IDS preference was also remarkably similar between the MA and the MLR: both data sources converged on an average effect size estimate of $d \approx 0.35$. The convergence of the estimate across the two sources of evidence in such a broad range of conditions can aid developmental scientists in sample size calculations for future studies of IDS preference (Lakens, 2022). For example, the effect size estimate of $d = 0.35$ implies that a sample at least as large as $N = 66$ infants is needed to have 80% or greater power to detect an IDS preference at an alpha level of .05 in a within-participant design using a paired-samples *t*-test. Our full dataset is also openly available, allowing researchers to account for potential sources of variability and tune their power estimates to specific methodological and modeling choices.

Why does IDS exert such an early, widespread effect on infants' preferential attention? One promising explanation posits that the engaging features of IDS reside in the mutual feedback loops between infant and caregiver, where infants' active participation and caregiver

responsiveness both contribute to the developmental process (Ko et al., 2016; Warlaumont et al., 2014). Given that adults use IDS as a consistent signal in addressing children during development, infants may start to associate the acoustic features of IDS with relevance and to recognize themselves as recipients of these salient utterances (Nencheva et al., 2021). This elevated attention to the speech stream, in turn, may drive the commonly observed language benefits of IDS during development (Golinkoff et al., 2015; Hartman et al., 2017; Peter et al., 2016).

At the same time, our exploratory analysis also found critical points on which the evidence from the MA and the MLR disagreed: infant age and experimental task showed distinct effects across the two sources. The different developmental trajectories of the IDS preference effect paint a complicated picture of the role of IDS during development. The linear increase with infant age in the MLR conforms to evidence that the IDS preference grows in response to experience with positive social interactions and increased participation in communicative exchanges (Ko et al., 2016; Warlaumont et al., 2014). On the other hand, the finding of stability across infant ages in the MA—which has also been previously reported in individual, smaller-scale studies in the literature (Newman & Hussain, 2006; Segal & Newman, 2015)—may indicate that IDS continues to be similarly relevant throughout early development.

The conflict in developmental trajectories in the MA and MLR may be driven by factors other than the underlying construct. For example, as discussed in the original ManyBabies 1 paper (The ManyBabies Consortium, 2020), the speech stimuli may have been best suited for the older age ranges in the study, or older infants may have exhibited more measurable behavioural responses. This would also accord with evidence that some acoustic characteristics of IDS change as children grow older (Cox et al., 2023). Conversely, in the MA, investigators had the freedom to tailor their stimuli and methods to the particular infant age investigated. One potential consequence of researchers tailoring methods to maximize effect sizes within the studied age range is that this practice may mask age-related changes in the strength of the IDS preference effect. This discrepancy between the results of the MA and MLR are not easily resolved. One way to improve our understanding of the developmental trajectories of the IDS preference would be to conduct more experiments on how infant looking time measures relate to their experience of the underlying construct (Kosie et al., 2023), and to use other higher-resolution non-behavioural measures to triangulate the effects that modulate infants' IDS preference (e.g., Nencheva et al., 2021).

The finding that experimental task produced diverging results across the MA and MLR again demonstrates limitations in the conclusions we can draw from each source on its own. For example, as discussed in the original paper (The ManyBabies Consortium, 2020), the finding of a stronger estimate in the MLR for studies using the HPP may be a function of the greater effort required on the part of the infant in the task, leading to stronger engagement and therefore to stronger effects. However, the MA did not demonstrate larger effect sizes for HPP methods, and at least numerically, the effect was in the opposite direction (see Figure 3). Smaller effect sizes for HPP compared to central fixation aligns with previous meta-analytic results in the infant literature (Bergmann et al., 2018). Taken at face value, these results call into question the generalizability of the result from the MLR. However, both the MLR and MA involved data from studies that self-selected the methodology employed to test the effect, severely limiting the causal inferences that can be drawn about the effect of methodology on IDS preference.[4] Future large-scale MLR studies may benefit from conducting random

[4] We should note that the goal of the MLR was not to replicate a single study, but rather to investigate how well the IDS preference generalized across different laboratories and methods. Because self-selection of methodologies likely varies systematically with other characteristics particular to each laboratory and study, we can at best make tentative conclusions about the effect of methodology on infants' IDS preference.

assignment of experimental methodology to participating labs; this experimental design would provide valuable information about the importance of methodological choices, the relation between MLRs and MAs, as well as how to interpret findings from infant studies more generally.

Our exploratory interaction analyses showed no robust differences in the effect of native language across the two sources of evidence. These results are consistent with the hypothesis that the main captivating features of IDS may reside in acoustic properties that are commonly attested across distinct languages (Cox et al., 2023; Hilton et al., 2022). We should note, however, that this result may have been driven in part by the unbalanced nature of the MA data, where only 5.4% of the effect sizes (vs. 54.9% in the MLR) included infant looking times to non-native speech stimuli. In the full sample of the original MLR (The ManyBabies Consortium, 2020), monolingual infants acquiring North American English had a stronger preference to attend to North American English IDS than monolinguals acquiring another language. The results here may thus be driven primarily by the imbalance in the MA effect sizes as well as the subsample characteristics of the MLR. This interpretation would also be in line with evidence from another recent MLR (Byers-Heinlein et al., 2021) showing that bilingual infants with a higher percentage of exposure to North American English had a stronger North American English IDS preference.

The overrepresentation of North American English in the MLR and especially in the MA is emblematic of the substantial language bias in developmental research (Christiansen et al., 2022; Kidd & Garcia, 2022; Kidd et al., 2023) and in IDS research in particular (Cox et al., 2023; Cristia, 2023; Ochs & Schieffelin, 1984). Oversampling from particular populations severely constrains our understanding of the global variability in the use of IDS across languages, dialects and cultures (Casillas et al., 2020; Cristia, 2023; Floccia et al., 2016), and this in turn limits our ability to construct generalizable theories about the features and functional relevance of IDS in different cultural settings. Based on the cumulative findings presented here (see Table 1), future research on IDS preference should focus on expanding language diversity both with respect to participants' language backgrounds and the speech stimuli tested, in order to evaluate the generalizability of the results to other sample characteristics (e.g., Tsui et al., 2023) and first languages (e.g., Soderstrom et al., in prep).

The complex interactions between sample characteristics in both the MA and MLR also highlight an important limitation in our conclusions: scarcity of available data on moderator interactions can hinder attribution of variation and accurate estimation in statistical models (Lipsey, 2003; Tipton et al., 2019). For example, all of the studies using artificial stimuli in the MA use a method that is neither HPP or central fixation, severely limiting the inferences we can draw about the effects of this stimulus type. This paper thus emphasizes the need for careful consideration and comprehensive assessment of moderator variables in future research to better understand and reconcile results across individual studies as well as MLRs and MAs (cf. Figure 1 in Section 4.2 of the Supplementary Materials). In the current context, theory-driven investigation of the extent to which the IDS preference effect is modulated by cross-linguistic variability in IDS features as well as differences in language exposure will be an important topic for future research.

Implications for the Relationship Between MAs and MLRs

Overall, both MLRs and MAs are useful techniques to combine and synthesize evidence from multiple studies. Each technique, however, has benefits and drawbacks. If used critically and with an understanding of its inherent limitations, MAs can serve as a crucial tool to assess the

progress of a field, to highlight its strengths and weaknesses, to provide methodological recommendations, and to offer directions for future research endeavors (Fusaroli et al., 2022; Nguyen et al., 2022). An inherent limitation of MAs, however, is that the data are filtered through the publication process. This process acts as a bias that selects for statistically significant findings, typically leading to an inflation of effect sizes in the MA (Kvarven et al., 2020; Lewis et al., 2022). Notably, however, our worst-case publication bias estimates for our MA were in fact *lower* than the MLR estimate, suggesting that estimates of the IDS preference phenomenon might not suffer from the same degree of publication bias as other phenomena in the developmental literature. MAs have also come under scrutiny for reasons beyond publication bias, including a lack of reproducibility and errors in the extraction of data (Maassen et al., 2020). MAs may be particularly susceptible to errors as they adopt any errors in the original studies (see e.g., Nuijten et al., 2016), combined with any new errors introduced by the MA. In the current paper, we found substantial errors in the original MA (Dunst et al., 2012), which changed the interpretation of some of the results (cf. Section 2.1 in the Supplementary Materials for a full list of revisions to the original MA; Section 5.3 for an overview of results across the original, revised and community-augmented datasets; and Section 6 for in-depth discussion of these discrepancies and our rationale in focusing on the community-augmented MA). In consideration of these limitations—including errors in reporting effect sizes, data curation errors, and omission of reported effect sizes—we call for higher standards in transparency of all steps of the meta-analytic process (Tsuji et al., 2014). These may fruitfully be pursued within already established open science initiatives for meta-scientific endeavours (e.g., MetaLab, https://metalab.stanford.edu/).

MLRs, on the other hand, can provide an estimate of the phenomenon of interest that is free from publication bias, but within a relatively restricted range of stimuli and methodological designs and with a very high cost in time and money. In the current context, individual labs were themselves allowed to select experimental methodology. Crucially, this limits the degree to which we can make causal inferences about the effect of methodology. One possible step that future MLRs could consider is randomly assigning participants to key moderators of interest (such as specific methodological choices). Manipulating a wider variety of moderators systematically would allow for stronger causal inferences and could lay the groundwork for a fuller understanding of the moderating role of design choices in the investigation of key phenomena.

Conclusions

In summary, we find robust evidence that IDS captivates infants' attention during development across two sources of evidence: a community-augmented MA and a MLR. Synthesizing the evidence from these two sources allowed us to show that IDS preference generalizes across a broad range of participants, ages, methods, and stimuli, albeit with substantial remaining uncertainty about how the magnitude of the IDS preference effect varies across key moderators. Many key questions about the IDS preference effect remain open. Evidence between the MLR and MA conflicts with respect to the developmental trajectory of IDS preference and the degree to which different methodologies elicit varying effect magnitudes. Overall, this study shows that MAs and MLRs provide distinct but complementary approaches to assessing phenomena and the factors that modulate them: MAs allow for estimating effects across heterogeneous design choices and populations in the extant literature, while MLRs offer an approach for large-scale, high-precision estimation of key effects within similar implementations and free from publication bias. Rather than considering either MAs or MLRs as the gold standard, this work demonstrates how integrating each of these two sources of evidence offers an attractive path forward for building cumulative evidence in psychological science.

ACKNOWLEDGMENTS

The funders had no role in the design, conduct, or reporting of this research. We would also like to thank the following research assistants: Lucy Anderson, Stephen Gilliat, Heewon Hwang, Sarah Kamhout, John Muldowney, and Taylor Orr.

FUNDING INFORMATION

This research was funded by SSHRC Partnership Development Grant GR019187 to MS. MBM was supported by NIH R01 LM013866-01. MZ was supported by a grant from the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number F32HD110174. Jessica Kosie was supported by NSF SBE Postdoctoral Research Fellowship 2004983 and NIH F32 F32HD103439.

AUTHOR CONTRIBUTIONS

The following lists each author's contribution to this paper based on CRediT (Contributor Roles Taxonomy). An overview of authorship contributions can be viewed here: https://docs.google.com/spreadsheets/d/1CQfw_ASSMT5boxpNGSfFmJvt8KQ0yiePgpwDEbRIrY0/edit?usp=sharing.

Martin Zettersten: Conceptualization, Data curation (lead), Data collection – coding papers (lead), Documentation, Formal analysis, Project administration, Software, Validation, Visualization, Writing – original draft (co-lead), Writing – review and editing (co-lead). Christopher Cox: Conceptualization, Data Curation, Formal analysis, Project administration, Software, Validation, Visualization, Writing - original draft (co-lead), Writing – review and editing (co-lead). Christina Bergmann: Conceptualization (lead), Data curation, Data collection – coding papers, Documentation, Formal analysis, Project administration, Software, Writing – original draft (co-lead), Writing – review and editing. Melanie Soderstrom: Writing – original draft, Writing – review and editing. Angeline Sin Mei Tsui: Conceptualization, Data collection – coding papers, Documentation, Writing – review and editing. Julien Mayor: Conceptualization, Writing – review and editing. Rebecca A. Lundwall: Data collection – coding papers, Resources, Writing – review and editing. Molly Lewis: Data curation, Formal analysis, Software. Jessica E. Kosie: Conceptualization, Data collection – coding papers, Data curation, Documentation, Writing – review and editing. Natalia Kartushina: Conceptualization, Data collection - coding papers, Data curation, Documentation, Writing – review and editing. Riccardo Fusaroli: Conceptualization, Software, Validation (lead), Writing – review and editing. Michael C. Frank: Conceptualization, Visualization, Writing – review and editing. Krista Byers-Heinlein: Conceptualization, Software, Visualization, Writing – original draft, Writing – review and editing. Alexis K. Black: Conceptualization, Data collection – coding papers, Data curation, Documentation, Writing – original draft. Maya B. Mathur: Conceptualization, Formal analysis (lead), Software (lead), Validation, Visualization (lead), Writing – original draft (co-lead), Writing – review and editing.

DATA AVAILABILITY STATEMENT

All code, materials, and data required to reproduce this research are publicly available and documented on OSF (https://osf.io/amj7u/) and GitHub (https://github.com/christinabergmann/IDSPreference_ManyBabiesMeta/).

REFERENCES

Anderson, L., Hwang, H., Kamhout, S., Giliat, S., Lundwall, R. A., Black, A., Kartushina, N., Kosie, J., Tsui, A., Zettersten, M., Bergmann, C., & The ManyBabies Consortium. (2021). *A fresh look at infant-directed speech preference through an updated meta-analysis*. Poster presented at the Biennial Meeting of the Society for Research in Child Development.

Bates, D., & Maechler, M. (2021). *Matrix: Sparse and dense matrix classes and methods* [Computer software manual]. Retrieved

from https://CRAN.R-project.org/package=Matrix (R package version 1.3-2).

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009. https://doi.org/10.1111/cdev.13079, PubMed: 29736962

Braginsky, M., Mathur, M., & VanderWeele, T. J. (2023). *PublicationBias: Sensitivity analysis for publication bias in meta-analyses* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=PublicationBias (R package version 2.4.0).

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., Durrant, S., Fennell, C. T., Fiévet, A.-C., Frank, M. C., Gampe, A., Gervain, J., Gonzales-Gomez, N., Hamlin, J. K., Havron, N., Hernik, M., Kerr, S., Killam, H., Klassen, K., … Wermelinger, S. (2021). A multi-lab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920974622, PubMed: 35821764

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a Tseltal Mayan village. *Child Development*, *91*(5), 1819–1835. https://doi.org/10.1111/cdev.13349, PubMed: 31891183

Christiansen, M. H., Kallens, P. C., & Trecca, F. (2022). Toward a comparative approach to language acquisition. *Current Directions in Psychological Science*, *31*(2), 131–138. https://doi.org/10.1177/09637214211049229

Coburn, K. M., & Vevea, J. L. (2019). *weightr: Estimating weight-function models for publication bias* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=weightr (R package version 2.0.2).

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, *61*(5), 1584–1595. https://doi.org/10.2307/1130766, PubMed: 2245748

Corker, K. S. (2022). Strengths and weaknesses of meta-analyses. In L. Jussim, S. Stevens, & J. Krosnick (Eds.), *Research integrity: Best practices for the social and behavioral sciences* (pp. 150–175). Oxford University Press. https://doi.org/10.1093/oso/9780190938550.003.0006

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & 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. https://doi.org/10.1038/s41562-022-01452-1, PubMed: 36192492

Cristia, A. (2023). A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. *Developmental Science*, *26*(1), e13265. https://doi.org/10.1111/desc.13265, PubMed: 35429106

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005, PubMed: 19285912

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., & Swinton, J. (2019). *xtable: Export tables to LaTeX or HTML* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=xtable (R package version 1.8-4).

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press. https://doi.org/10.1017/CBO9780511802843

Dowle, M., & Srinivasan, A. (2020). *data.table: Extension of 'data.frame'* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=data.table (R package version 1.13.2).

Dunst, C. J., Gorman, E., & Hamby, D. W. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning Reviews*, *5*(1), 1–13.

Eaves, B. S., Feldman, N. H., Griffiths, T. L., & Shafto, P. (2016). Infant-directed speech is consistent with teaching. *Psychological Review*, *123*(6), 758–771. https://doi.org/10.1037/rev0000031, PubMed: 27088361

Ebersole, C. R., Andrighetto, L., Casini, E., Chiorri, C., Dalla Rosa, A., Domaneschi, F., Ferguson, I. R., Fryberger, E., Giacomantonio, M., Grahe, J. E., Joy-Gaba, J. A., Langford, E. V., Nichols, A. L., Panno, A., Parks, K. P., Preti, E., Richetin, J., & Vianello, M. (2020). Many Labs 5: Registered replication of Payne, Burkley, and Stokes (2008), Study 4. *Advances in Methods and Practices in Psychological Science*, *3*(3), 387–393. https://doi.org/10.1177/2515245919885609

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*, 517. https://doi.org/10.1037//0003-066X.33.5.517.a

Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, *10*(3), 279–293. https://doi.org/10.1016/0163-6383(87)90017-8

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501. https://doi.org/10.1017/S0305000900010679, PubMed: 2808569

Fisher, Z., Tipton, E., & Zhipeng, H. (2017). *robumeta: Robust variance meta-regression* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=robumeta (R package version 2.0).

Floccia, C., Keren-Portnoy, T., DePaolis, R., Duffy, H., Delle Luche, C., Durrant, S., White, L., Goslin, J., & Vihman, M. (2016). British english infants segment words only with exaggerated infant-directed speech stimuli. *Cognition*, *148*, 1–9. https://doi.org/10.1016/j.cognition.2015.12.004, PubMed: 26707426

Fusaroli, R., Grossman, R., Bilenberg, N., Cantio, C., Jepsen, J. R. M., & Weed, E. (2022). Toward a cumulative science of vocal markers of autism: A cross-linguistic meta-analysis-based investigation of acoustic markers in american and danish autistic children. *Autism Research*, *15*(4), 653–664. https://doi.org/10.1002/aur.2661, PubMed: 34957701

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)Talk to me: The social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344. https://doi.org/10.1177/0963721415595345

Hartman, K. M., Ratner, N. B., & Newman, R. S. (2017). Infant-directed speech (IDS) vowel clarity and child language outcomes. *Journal of Child Language*, *44*(5), 1140–1162. https://doi.org/10.1017/S0305000916000520, PubMed: 27978860

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5, PubMed: 26056092

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., Knox, D., Glowacki, L., Alemu, E., Galbarczyk, A., Jasienska, G., Ross, C. T., Neff, M. B., Martin, A., Cirelli, L. K., Trehub, S. E., Song, J., Kim, M., ... Mehr, S. A. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*, *6*(11), 1545–1556. https://doi.org/10.1038/s41562-022-01410-x, PubMed: 35851843

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., ... Coles, N. A. (2021). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour*, *5*, 159–169. https://doi.org/10.1038/s41562-020-01007-2, PubMed: 33398150

Kalashnikova, M., & Burnham, D. (2018). Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions. *Journal of Child Language*, *45*(5), 1035–1053. https://doi.org/10.1017/S0305000917000629, PubMed: 29502549

Kaplan, J. (2020). *fastdummies: Fast creation of dummy (binary) columns and rows from categorical variables* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=fastDummies (R package version 1.6.3).

Kidd, E., Christiansen, M., Majid, A., Bornkessel-Schlesewsky, I., Schlesewsky, M., & Evans, N. (2023). Harnessing linguistic diversity for theories of language and mind. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th annual conference of the Cognitive Science Society* (pp. 18–19). Cognitive Science Society.

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735. https://doi.org/10.1177/01427237211066405

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., Ahn, P. H., Brady, A. J., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J., Cromar, R., Gardiner, G., Gosnell, C., Grahe, J., Hall, C., Howard, I., ... Ratliff, K. (2019). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *PsyArXiv*. https://doi.org/10.31234/osf.io/vef2c

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., ... Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., ... Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Ko, E.-S., Seidl, A., Cristia, A., Reimchen, M., & Soderstrom, M. (2016). Entrainment of prosody in the interaction of mothers with their young children. *Journal of Child Language*, *43*(2), 284–309. https://doi.org/10.1017/S0305000915000203, PubMed: 26036694

Kosie, J., Zettersten, M., Abu-Zhaya, R., Amso, D., Babineau, M., Baumgartner, H. A., Bazhydai, M., Belia, M., Benavides-Varela, S., Bergmann, C., Berteletti, I., Black, A. K., Borges, P., Borovsky, A., Byers-Heinlein, K., Cabrera, L., Calignano, G., Cao, A., Chijiiwa, H., ... Lew-Williams, C. (2023). Manybabies 5: A large-scale investigation of the proposed shift from familiarity preference to novelty preference in infant looking time. *PsyArxiv*. https://doi.org/10.31234/osf.io/ck3vd

Kuhn, M., Jackson, S., & Cimentada, J. (2020). *corrr: Correlations in R* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=corrr (R package version 0.4.3).

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z, PubMed: 31873200

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*(1), 24. https://doi.org/10.1186/s40359-016-0126-3, PubMed: 27241618

Lewis, M., Mathur, M., VanderWeele, T., & Frank, M. C. (2022). The puzzling relationship between multi-laboratory replications and meta-analyses of the published literature. *Royal Society Open Science*, *9*(2), 211499. https://doi.org/10.1098/rsos.211499, PubMed: 35223059

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *ANNALS of the American Academy of Political and Social Science*, *587*(1), 69–81. https://doi.org/10.1177/0002716202250791

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLoS One*, *15*(5), e0233107. https://doi.org/10.1371/journal.pone.0233107, PubMed: 32459806

Makowski, D., Lüdecke, D., Patil, I., Thériault, R., Ben-Shachar, M., & Wiernik, B. (2023). Automated results reporting as a practical tool to improve reproducibility and methodological best practices adoption. *CRAN*. Retrieved from https://github.com/easystats/report.

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335, PubMed: 22853650

Mathur, M. B., & VanderWeele, T. J. (2019). New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine*, *38*(8), 1336–1342. https://doi.org/10.1002/sim.8057, PubMed: 30513552

Mathur, M. B., & VanderWeele, T. J. (2020a). Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. *Epidemiology*, *31*(3), 356–358. https://doi.org/10.1097/EDE.0000000000001180, PubMed: 32141922

Mathur, M. B., & VanderWeele, T. J. (2020b). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society: Series C*, *69*(5), 1091–1119. https://doi.org/10.1111/rssc.12440, PubMed: 33132447

Mathur, M. B., & VanderWeele, T. J. (2021a). Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Research Synthesis Methods*, *12*(2), 176–191. https://doi.org/10.1002/jrsm.1464, PubMed: 33108053

Mathur, M. B., & VanderWeele, T. J. (2021b). Meta-regression methods to characterize evidence strength using meaningful-effect percentages conditional on study characteristics. *Research Synthesis Methods*, *12*(6), 731–749. https://doi.org/10.1002/jrsm.1504, PubMed: 34196505

McShane, B. B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, *112*(519), 885–895. https://doi.org/10.1080/01621459.2017.1289846

Müller, K. (2020). *here: A simpler way to find your files* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=here (R package version 1.0.1).

Müller, K., & Wickham, H. (2021). *tibble: Simple data frames* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tibble (R package version 3.1.1).

Nencheva, M. L., Piazza, E. A., & Lew-Williams, C. (2021). The moment-to-moment pitch dynamics of child-directed speech shape toddlers' attention and learning. *Developmental Science*, *24*(1), e12997. https://doi.org/10.1111/desc.12997, PubMed: 32441385

Newman, R. S., & Hussain, I. (2006). Changes in preference for infant-directed speech in low and moderate noise by 4.5- to 13-month-olds. *Infancy*, *10*(1), 61–76. https://doi.org/10.1207/s15327078in1001_4, PubMed: 33412673

Nguyen, V., Versyp, O., Cox, C., & Fusaroli, R. (2022). A systematic review and bayesian meta-analysis of the development of turn taking in adult-child vocal interactions. *Child Development*, *93*(4), 1181–1200. https://doi.org/10.1111/cdev.13754, PubMed: 35305028

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2, PubMed: 26497820

Ochs, E., & Schieffelin, B. B. (1984). Language acquisition and socialization: Three developmental stories and their implications. In R. A. Shweder & R. A. LeVine (Eds.), *Culture theory: Essays on mind, self, and emotion* (pp. 276–320). Cambridge University Press.

Pegg, J. E., Werker, J. F., & McLeod, P. J. (1992). Preference for infant-directed over adult-directed speech: Evidence from 7-week-old infants. *Infant Behavior and Development*, *15*(3), 325–345. https://doi.org/10.1016/0163-6383(92)80003-D

Peter, V., Kalashnikova, M., Santos, A., & Burnham, D. (2016). Mature neural responses to infant-directed speech but not adult-directed speech in pre-verbal infants. *Scientific Reports*, *6*(1), 34273,https://doi.org/10.1038/srep34273, PubMed: 27677352

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software manual]. Retrieved from https://www.R-project.org/.

Rich, B. (2021). *table1: Tables of descriptive statistics in HTML* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=table1 (R package version 1.4).

Segal, J., & Newman, R. S. (2015). Infant preferences for structural and prosodic properties of infant-directed speech in the second year of life. *Infancy*, *20*(3), 339–351. https://doi.org/10.1111/infa.12077

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: A best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology*, *70*, 747–770. https://doi.org/10.1146/annurev-psych-010418-102803, PubMed: 30089228

Simonsohn, U., Simmons, J., & Nelson, L. D. (2022). Above averaging in literature reviews. *Nature Reviews Psychology*, *1*, 551–552. https://doi.org/10.1038/s44159-022-00101-8

Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? *Infancy*, *3*(3), 365–394. https://doi.org/10.1207/S15327078IN0303_5, PubMed: 33451217

Snow, C. E., & Ferguson, C. A. (Eds.) (1977). *Talking to children: Language input and acquisition*. Cambridge University Press.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Soderstrom, M., Junge, C., Kartushina, N., Soley, G., Mayor, J., Durier, V., Barbu, S., Oceláková, Z., Chladkova, K., Smolík, F., & Fikkert, P. *ManyBabies1 native language follow-up: Preference for infant-directed speech across languages*. Retrieved 2024-01-23, from https://osf.io/9j87t/ (Manuscript in preparation).

Stanley, T. D. (2001). Wheat from chaff: Meta-analysis as quantitative literature review. *Journal of Economic Perspectives*, *15*(3), 131–150. https://doi.org/10.1257/jep.15.3.131

Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language*, *10*(1), 1–15. https://doi.org/10.1017/S0305000900005092, PubMed: 6841483

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ*, *323*(7304), 101–105. https://doi.org/10.1136/bmj.323.7304.101, PubMed: 11451790

The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2019). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*(2), 161–179. https://doi.org/10.1002/jrsm.1338, PubMed: 30589224

Tsui, A., Carstensen, A., Kachergis, G., Abubakar, A., Asnake, M., Barry, O., Basnight-Brown, D. M., Bentu, D., Bergmann, C., Dami, E. B., Boll-Avetisyan, N., de Jongh, M., Diot, Y., Duah, R. A., Herrmann, E., Jang, C., Kizito, S., Lamba, T., Maliwichi-Senganimalunje, L., ... Frank, M. C. (2023). *Exploring variation in infants' preference for infant-directed speech: Evidence from a multi-site study in Africa*. Retrieved 2024-01-23, from https://osf.io/fqp4b. https://doi.org/10.17605/OSF.IO/JGR79

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science*, *9*(6), 661–665. https://doi.org/10.1177/1745691614552498, PubMed: 26186116

Ushey, K., & Wickham, H. (2021). *renv: Project environments* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=renv (R package version 0.13.2).

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*(3), 419–435. https://doi.org/10.1007/BF02294384

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., Franchin, L., Frank, M. C., Geraci, A., Hamlin, J. K., Kaldy, Z., Kulke, K., Laverty, C., Lew-Williams, C., Mateu, V., Mayor, J., Moreau, D., Nomikou, I., Schuwerk, T., ... Zettersten, M. (2022). Improving the generalizability of infant psychological research: The ManyBabies model. *Behavioral and Brain Sciences*, *45*, e35. https://doi.org/10.1017/S0140525X21000455, PubMed: 35139960

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. *Psychological Science*, *25*(7), 1314–1324.

https://doi.org/10.1177/0956797614531023, PubMed: 24840717

Werker, J. F., & McLeod, P. J. (1989). Infant preference for both male and female infant-directed talk: A developmental study of attentional and affective responsiveness. *Canadian Journal of Psychology, 43*(2), 230–246. https://doi.org/10.1037/h0084224, PubMed: 2486497

Wickham, H. (2011). testthat: Get started with testing. *R Journal, 3*(1), 5–10. https://doi.org/10.32614/RJ-2011-002

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H. (2019). *stringr: Simple, consistent wrappers for common string operations* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=stringr (R package version 1.4.0).

Wickham, H. (2021a). *forcats: Tools for working with categorical variables (factors)* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=forcats (R package version 0.5.1).

Wickham, H. (2021b). *tidyr: Tidy messy data* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tidyr (R package version 1.1.3).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2021). *dplyr: A grammar of data manipulation* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dplyr (R package version 1.0.5).

Wickham, H., & Henry, L. (2020). *purrr: Functional programming tools* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=purrr (R package version 0.3.4).

Wickham, H., & Hester, J. (2020). *readr: Read rectangular text data* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=readr (R package version 1.4.0).

Wickham, H., Pedersen, T. L., & Seidel, D. (2020). *scales: Scale functions for visualization* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=scales (R package version 1.1.1).

Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible research* (pp. 3–31). Chapman and Hall/CRC. https://doi.org/10.1201/9781315373461-1

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences, 45*, e1. https://doi.org/10.1017/S0140525X20001685, PubMed: 33342451

Yoshida, K., & Bartel, A. (2020). *tableone: Create 'table 1' to describe baseline characteristics with or without propensity score weights* [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=tableone (R package version 0.12.0).