# 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[10], Alexis K. Black[11], and Maya B. Mathur[12]

[1]Department of Psychology, Princeton University
[2]Department of Cognitive Science and Interacting Minds Center, Aarhus University
[3]Max Planck Institute for Psycholinguistics
[4]Department of Psychology, Stanford University
[5]Department of Psychology, University of Manitoba
[6]Department of Psychology, University of Oslo
[7]Psychology Department and Neuroscience Center, Brigham Young University
[8]Department of Psychology/Social and Decision Sciences, Carnegie Mellon University
[10]Department of Psychology, Concordia University
[11]School of Audiology and Speech Sciences, University of British Columbia
[12]Quantitative Sciences Unit, Stanford University

*Authors marked with * share joint first-authorship. Author order outside of the first three and final positions is reverse alphabetical.

†Correspondence to: Martin Zettersten (martincz@princeton.edu), Department of Psychology, South Dr, Princeton, NJ, 08540.

## 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 ($d = 0.35$ for both estimates) 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.

# 1. Introduction

Across many cultures, adults adjust the way they speak with infants compared to how they speak with other adults (Hilton et al., 2022; Cox et al., 2022; Fernald et al., 1989). 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 (Stern et al., 1983; Hilton et al., 2022; Singh et al., 2002; Kalashnikova & Burnham, 2018). 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; Pegg et al., 1992; Werker & McLeod, 1989; Fernald & Kuhl, 1987) 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 (Fernald et al., 1989; Eaves et al., 2016; Peter et al., 2016; Soderstrom, 2007; Hartman et al., 2017; Cox et al., 2022; Snow & Ferguson, 1977; Csibra & Gergely, 2009).

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 (Dunst et al., 2012; Anderson et al., 2021) 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 (http://metalab.stanford.edu; Bergmann et al., 2018; Anderson et al., 2021). 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, 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 (Lakens et al., 2016; Stanley, 2001; Siddaway et al., 2019; Corker, 2022). 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 (Mathur & VanderWeele, 2021a; Sterne et al., 2001; McShane & Gal, 2017; Masicampo & Lalande, 2012). 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., Klein et al., 2014, 2018, 2019; Ebersole et al., 2016, 2020; The ManyBabies Consortium, 2020; Jones et al., 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 (Yarkoni, 2022; Visser et al., 2022).

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 17 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, and so unmeasured moderators may be important. Given the limitations of MAs and MLRs considered alone – and resulting discrepancies in conclusions – a promising approach to understanding a key phenomenon of interest may therefore be 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 evidence 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 thought to substantially impact IDS preference. 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

# 2. Methods

All confirmatory analyses were preregistered prior to data analysis at `https://osf.io/scg9z?view_only=7fd9e41122e042cfa998e50cf0336572`. 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).

## 2.1. Meta-analysis

### 2.1.1 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 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 additional 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).

### 2.1.2 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).

### 2.1.3 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 (http://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.

well as a discussion of discrepancies, see Supplementary Materials (Section 5 and Section 6).

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. 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.

## 2.2. 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 Section 3.1 and 3.2 of the Supplementary Materials.

## 2.3. 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.

## 2.4. Statistical analyses

### 2.4.1 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 estimated the percentage of population effects[2] in each source that were positive, representing any preference for IDS regardless of magnitude (Mathur & VanderWeele, 2020b, 2019). 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, 2020b, 2019) (in the predicted direction, i.e. showing an IDS preference).

### 2.4.2 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.

[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 & Seidel, 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, 2021), here (Müller, 2020), tibble (Müller & Wickham, 2021), purrr (Henry & Wickham, 2020), report (Makowski et al., 2021), data.table (Dowle & Srinivasan, 2020), corrr (Kuhn et al., 2020), PublicationBias (Mathur & VanderWeele, 2020a), metafor (Viechtbauer, 2010), tidyverse (Wickham et al., 2019), knitr (Xie, 2014), and robumeta (Fisher et al., 2017)

**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%) |
| 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%) |

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 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.

### 2.4.3 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, 2020c), and the significance funnel plot (Mathur & VanderWeele, 2020c). 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; McShane & Gal, 2017; Masicampo & Lalande, 2012). 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, 2020c). 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, 2020c). 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.

# 3. Results

## 3.1. 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 $\widehat{\tau} = 0.31$). 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 ($\widehat{\tau} = 0.11$) compared to the MA (cf., similar results when applying more stringent participant inclusion criteria on the MLR in Section 5.3.6 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 5** in the Supplementary Materials Section 5.3.3 for a visualization of the estimated densities of population effects, illustrating the greater heterogeneity of the MA as compared to the MLR).

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).

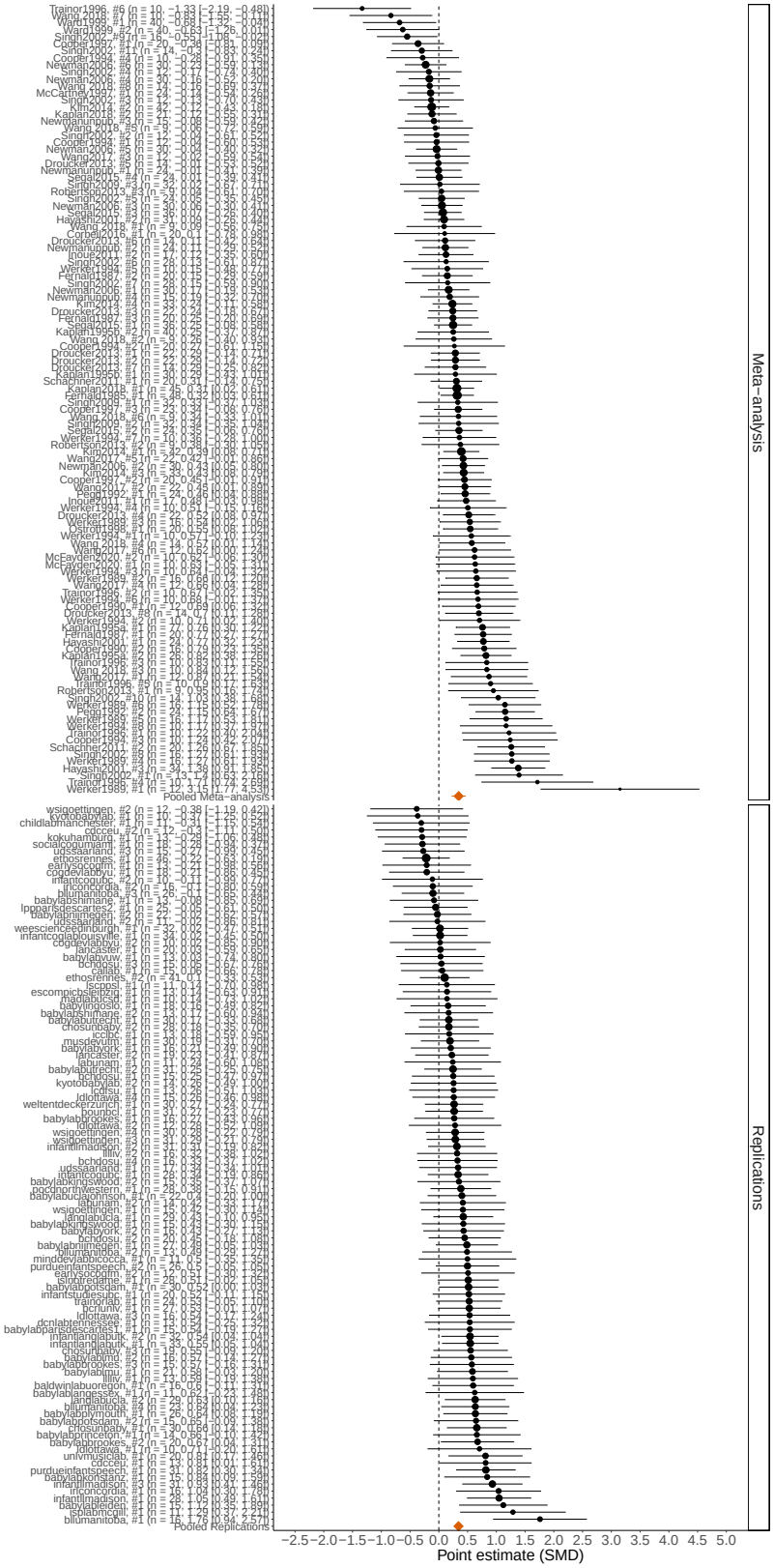


**Figure 1:** *Forest plot of studies' point estimates and 95% confidence intervals in the MA (top panel) and MLR (bottom panel). Orange diamonds: pooled estimates within each source. Dashed vertical line: null.*

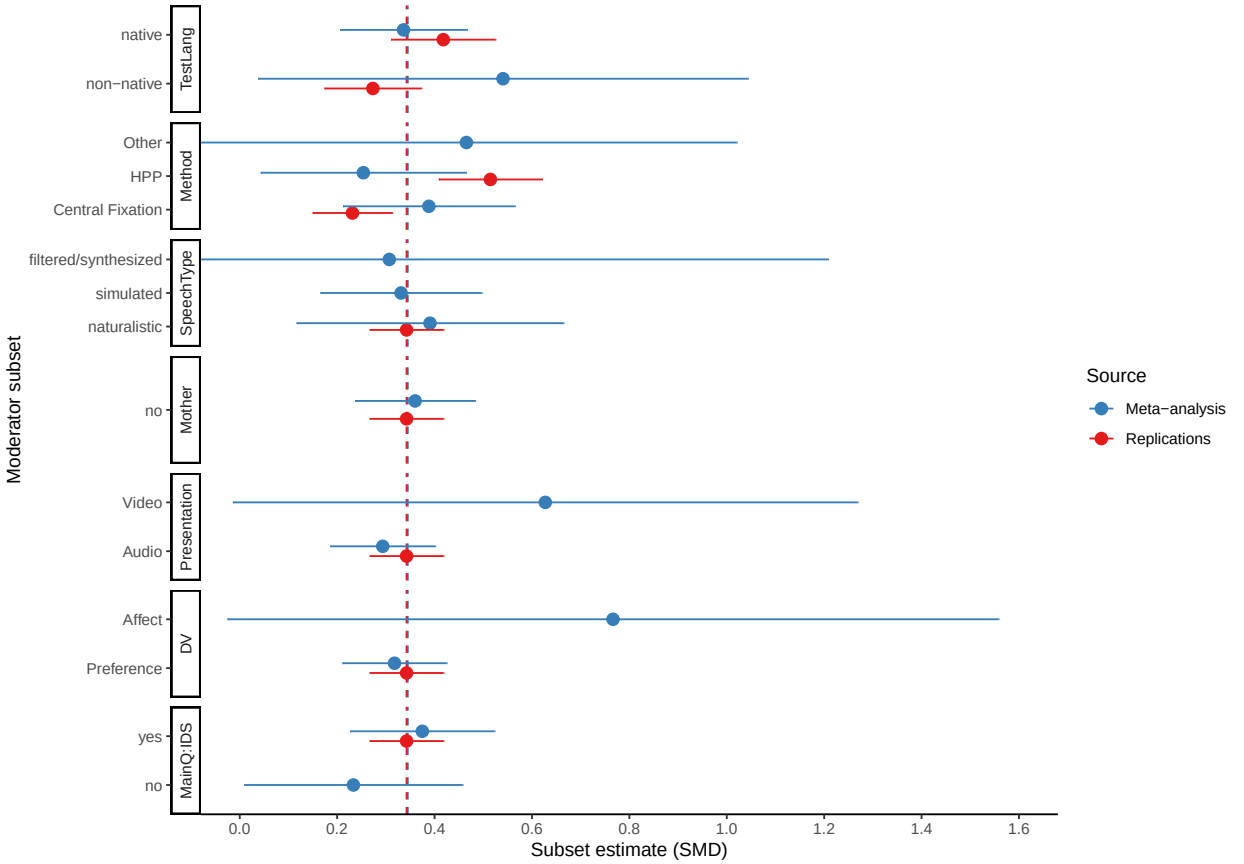


**Figure 2:** *Forest plot showing, for each categorical candidate moderator, the pooled point estimates for the subset of studies in the MA and in the MLR, respectively, with a given level of the moderator (including only levels with at least 5 observations). Error bars are 95% confidence intervals. 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.*

| Statistical measure | Unadjusted model | Moderated model |
|---|---|---|
| $\widehat{\mu}$ in MA | 0.34 [0.22, 0.46] | 0.32 [0.16, 0.47] |
| $\widehat{\mu}$ in MLR | 0.34 [0.27, 0.42] | 0.35 [0.22, 0.47] |
| $\widehat{\mu}$ 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] |

**Table 2:** *$\widehat{\mu}$: 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 $\widehat{\mu}$ measures (Hedges et al., 2010) and for differences in $\widehat{\mu}$ between sources and are bootstrapped for the percentage measures and for all cross-model comparisons (Mathur & VanderWeele, 2020b, 2021b). Confidence intervals are omitted when they were not statistically estimable (i.e., for percentage estimates that were very close to 0% or 100%).*

## 3.2. 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 $\widehat{\tau}_{\text{unadjusted}} = 0.27$). 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 unadjusted model ($\widehat{\tau}_{\text{mod}} = 0.30$). 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.

| 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 |

**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.*

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 ($b = 0.00$ [$-0.02$, 0.02]; $p$ =0.82; Figure 4A). Second, we found a significant interaction 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 4B). 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 ($\widehat{\tau}_{\text{mod}} = 0.27$) 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 5.3.7 of the Supplementary Materials).

| 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) | 0 | [-0.17, 0.17] | 0.973 |
| Method (HPP vs. Other) | 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 |

**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.*

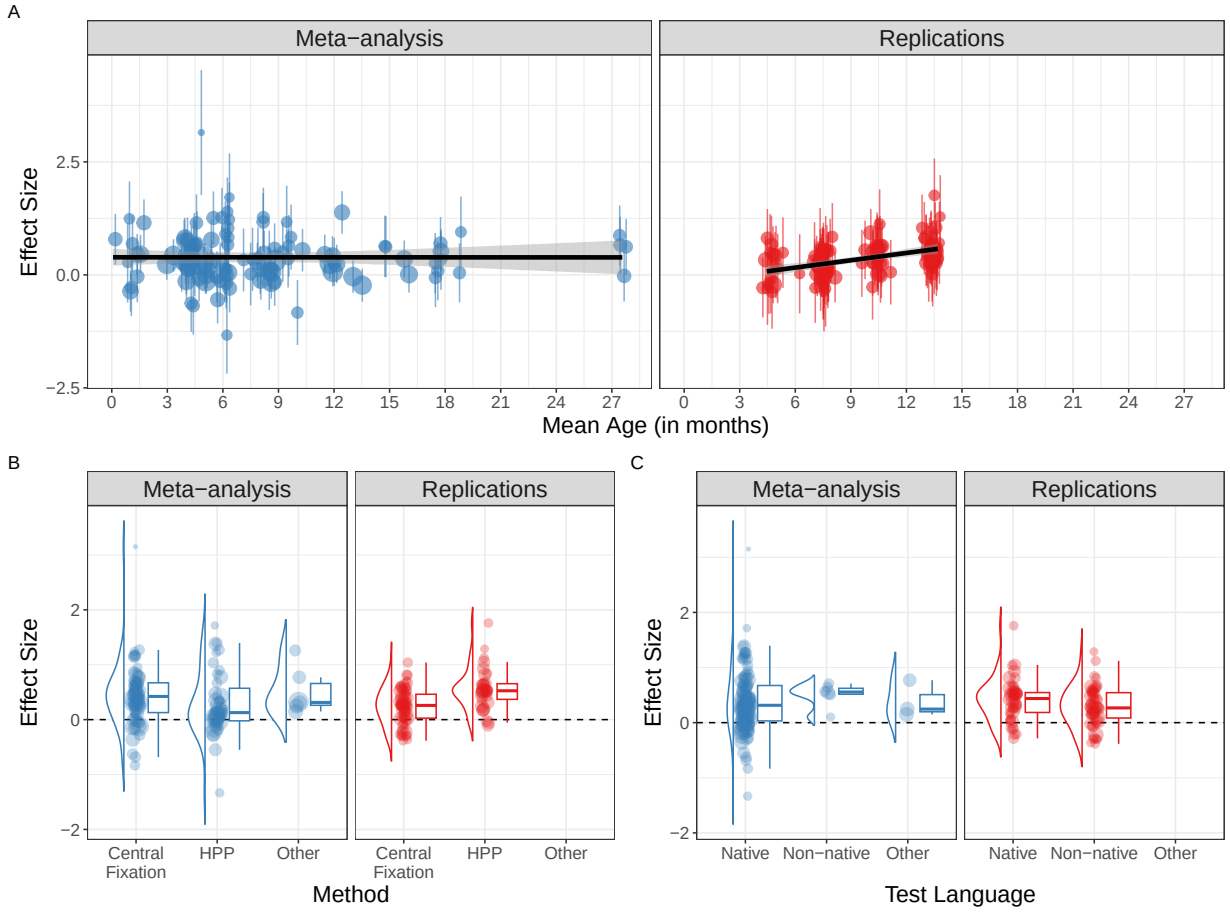


**Figure 3:** *Overview of the distribution of effect sizes in the meta-analysis (MA) and replications (MLR) for three key moderators: infant age (A), method (B), and test language (C). In (A), the black line represents a linear fit through the effect sizes for each source and error bars for individual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

### 3.3. 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 5.3.5 of the Supplementary Materials for additional analyses of publication bias).

## 4. 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 large-scale 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.

### 4.1. 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; Pegg et al., 1992; Werker & McLeod, 1989; Fernald & Kuhl, 1987). 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 (Warlaumont et al., 2014; Ko et al., 2016). 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 (Warlaumont et al., 2014; Ko et al., 2016). 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 (Segal & Newman, 2015; Newman & Hussain, 2006) – 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., 2022). 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, Zettersten, 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).

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., 2022; 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. These complex interactions between sample characteristics in both the MA and MLR 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 highlights 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.

## 4.2. Implications for the relationship between MAs and MLRs

The finding that experimental methodology produced diverging results across the MA and MLR again highlights important 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, leading to stronger engagement in the task 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 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.

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

[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.

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). In consideration of these limitations - namely biased reporting, researcher degrees of freedom and data curation errors – 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, http://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.

## 4.3. 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.

## Reproducibility Statement

All code, materials, and data required to reproduce this research are publicly available and documented (https://osf.io/amj7u/?view_only=379794009374448b8c7dd735a70a3198).

## Acknowledgments

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. 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.

## 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 (human), 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.

# References

Anderson, L., Hwang, H., Kamhout, S., Giliat, S., Lundwall, R., Black, A., ... Bergmann, C. (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.

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., & Black, A. K. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920974622.

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*, 1584–1595.

Corker, K. S. (2022). Strengths and weaknesses of meta-analyses. In L. Jussim, S. Stevens, & J. Krosnick (Eds.), *Research integrity in the behavioral sciences.* Oxford: Oxford University Press.

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2022). A systematic review and bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, *7*, 114-133. doi: 10.1038/s41562-022-01452-1

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in cognitive sciences*, *13*(4), 148–153. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1364661309000473`

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 applications.* Cambridge: Cambridge University Press. Retrieved from `http://statwww.epfl.ch/davison/BMA/` (ISBN 0-521-57391-2)

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., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *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. doi: 10.1037/rev0000031

Ebersole, C. R., Andrighetto, L., Casini, E., Chiorri, C., Dalla Rosa, A., Domaneschi, F., ... others (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. doi: https://doi.org/10.1177/2515245919885609

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... others (2016). Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. doi: https://doi.org/10.1016/j.jesp.2015.10.012

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*, 517–517. (Place: US Publisher: American Psychological Association) doi: 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.

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.

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)

Fusaroli, R., Grossman, R., Bilenberg, N., Cantio, C., Jepsen, J. R., Møllegaard, & 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.

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). The moment-to-moment pitch dynamics of child-directed speech shape toddlers' attention and learning. *Current Directions in Psychological Science*, *24*(5), 339-344.

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*, 1140–1162.

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.

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

Hilton, C. B., Moser, C. J., & Bertolo, M. e. a. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*, *6*, 1545-1556. doi: 10.1038/s41562-022-01410-x

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. Retrieved from https://www.jstatsoft.org/v42/i08/

Jones, B., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., ... Peters, K. O. (2020). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour*.

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.

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)

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., ... et al. (2019, Dec). *Many labs 4: Failure to replicate mortality salience effect with and without original author involvement.* PsyArXiv. Retrieved from psyarxiv.com/vef2c doi: 10.31234/osf.io/vef2c

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... others (2014). Investigating variation in replicability. *Social psychology*. doi: https://doi.org/10.1027/1864-9335/a000178.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Reginald B. Adams, J., Alper, S., ... 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. doi: 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.

Kosie, J., Zettersten, M., & The ManyBabies 5 Consortium. (2023). Manybabies 5: A large-scale investigation of the proposed shift from familiarity preference to novelty preference in infant looking time. *PsyArxiv*. Retrieved from https://psyarxiv.com/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. doi: https://doi.org/10.1038/s41562-019-0787-z

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

Lewis, M., Mathur, M., VanderWeele, T., & Frank, M. C. (2022). The puzzling relationship between multi-lab replications and meta-analyses of the rest of the literature. *Royal Society Open Science*, *9*(2), 211499.

Lipsey, M. W. (2003). Those confounded moderators in meta-analysis: Good, bad, and ugly. *ANNALS of the American Academy of Political and Social Science*, *897*(1), 69-81.

Maassen, E., van Assen, M. A., 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.

Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2021). 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., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279.

Mathur, M. B., & VanderWeele, T. J. (2019). New metrics for meta-analyses of heterogeneous effects. *Statistics in Medicine*, *38*(8), 1336–1342.

Mathur, M. B., & VanderWeele, T. J. (2020a). Publicationbias: Sensitivity analysis for publication bias in meta-analyses [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=PublicationBias` (R package version 2.2.0)

Mathur, M. B., & VanderWeele, T. J. (2020b). Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. *Epidemiology*, *31*(3), 356–358.

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

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.

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.

McShane, B. B., & Gal, D. (2017). Statistical significance and the dichotomization of evidence. *Journal of the American Statistical Association*, *112*(519), 885–895.

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.

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.

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.

Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, *48*(4), 1205–1226.

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.

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.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. 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.

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*(1), 747-770. (PMID: 30089228) doi: 10.1146/annurev-psych-010418-102803

Simonsohn, U., Simmons, J., & Nelson, L. D. (2022, October). Above averaging in literature reviews. *Nature Reviews Psychology*, *1*(10), 551–552. Retrieved 2023-05-29, from https://www.nature.com/articles/s44159-022-00101-8 (Number: 10 Publisher: Nature Publishing Group) doi: 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*, 365–394.

Snow, C. E., & Ferguson, C. A. (Eds.). (1977). *Talking to Children: Language Input and Acquisition*. Cambridge: 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.

Stanley, T. D. (2001, September). Wheat from chaff: Meta-analysis as quantitative literature review. *Journal of Economic Perspectives*, *15*(3), 131-150. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/jep.15.3.131` doi: 10.1257/jep.15.3.131

Stern, D. N., Spieker, S., Barnett, R., & MacKain, K. (1983). The prosody of maternal speech: infant age and context related changes. *Journal of Child Language*, *10*, 1-15.

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Investigating and dealing with publication and other biases in meta-analysis. *BMJ*, *323*(7304), 101-105.

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. doi: 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.

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

Ushey, K. (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. Retrieved from `https://www.jstatsoft.org/v36/i03/`

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., ... Zettersten, M. (2022). Improving the generalizability of infant psychological research: The manybabies model. *Behavioral and Brain Sciences*, *45*. doi: 10.1017/S0140525X21000455

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*, 1314–1324.

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.

Wickham, H. (2011). testthat: Get started with testing. *The R Journal*, *3*, 5–10. Retrieved from https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

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., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (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., & 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., & 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* (V. Stodden, F. Leisch, & R. D. Peng, Eds.). Chapman and Hall/CRC. Retrieved from http://www.crcpress.com/product/isbn/9781466561595 (ISBN 978-1466561595)

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, 1-78. doi: 10.1017/S0140525X20001685

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)

# Supplementary Materials: Evidence for infant-directed speech preference is consistent across large-scale, multi-site replication and meta-analysis

## Contents

# 1. Preregistration

## 1.1. Preregistration Approach

All confirmatory analyses were preregistered prior to data analysis (`https://osf.io/scg9z?view_only=7fd9e41122e042cfa998e50cf0336572`). Given that all analyses involved pre-existing datasets, we outline how the team approached accessing and curating the data in relation to the development of the preregistration protocol below. We wrote the preregistration protocol after we had accessed the original MA dataset (a digitized version of information in the Appendices of Dunst et al. (2012)), assembled the replication data (via a public Github repository at https://github.com/manybabies/mb1-analysis-public), and conducted basic cleaning on both datasets but before conducting any analyses relevant to the preregistered research questions. During the process of developing and planning statistical analyses, the statistician (MBM) was provided with only a "dummy" version of the combined dataset (comprising both the MA and MLR data) in which the point estimates and their variances had been randomly permuted across the two sources (i.e. the MA and the MLR). All authors co-developed the preregistration protocol, some of whom had access to the veridical dataset during protocol development but who had not conducted any of the planned analyses. Note that coauthors were aware of the main results of the original MA and the MLR as they were reported in the published reports (Dunst et al., 2012; The ManyBabies Consortium, 2020).

## 1.2. Changes and additions to preregistered protocol

Here, we describe and justify deviations from our preregistered protocol.

1. We focused all analyses on a revised and expanded community-augmented meta-analysis (`https://langcog.github.io/metalab`) instead of the original (Dunst et al., 2012) meta-analysis. Our preregistered plan was to conduct the meta-analysis by transcribing the dataset from (Dunst et al., 2012), with several additional moderator variables coded. However, over the course of the project, we revised and expanded the meta-analysis in two ways. First, we revised substantial issues discovered in the Dunst et al. (2012) meta-analysis (see Section 2.1 for details). Second, the meta-analysis was substantially augmented by the metalab community of infant researchers, leading to a significantly expended meta-analysis (30 papers, 112 effect sizes). We ultimately chose to focus our primary analyses on the community-augmented meta-analysis as the most comprehensive, accurate meta-analytic dataset on IDS preference available. All analyses with the original and revised Dunst et al. (2012) datasets are reported below (Section 5) and discrepancies between the datasets are discussed in more depth (Section 6).

2. Our preregistration specified that we would estimate the percentages of positive effects and of effects stronger than $SMD = 0.20$ for each source by using a single meta-regression model (Mathur & VanderWeele, 2021). However, the heterogeneity estimate in such a model is an average over the two sources, and it became apparent during data analysis that the MA showed considerably more heterogeneity than the MLR. We

therefore estimated the percentage metrics for each source by fitting separate meta-analysis or meta-regression models, as described in the main text. For the same reason, we omitted an analysis we had planned in which we would have estimated the difference in the heterogeneity estimates from the unadjusted model versus the moderated model, each containing data from both sources.

3. We added additional exploratory analyses to investigate the interaction between study-level moderators and the two data sources. These exploratory analyses were added to enable estimation of discrepancies between the two sources in terms of the effect of the three predictors (i.e., infant age, test language and method) on IDS preference. 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. We describe this model as exploratory throughout the main manuscript.

## 2. Revising and augmenting the meta-analysis

### 2.1. Revisions to Dunst et al. (2012)

While coding additional moderators for the studies included in a previous meta-analysis of infant-directed speech preference (Dunst et al., 2012), we encountered substantial issues with the results and study classifications reported in the original meta-analysis. The main issues we identified are as follows.

- **incorrectly reported effect sizes.** The effect sizes for some papers were inconsistent with the effect sizes we determined based on reported statistics and figures in the original papers. In some cases, we could identify the source of the error (e.g., incorrectly treating the condition manipulation as between-subjects rather than within-subjects when computing effect size), while in other cases we were unable to trace the source of the incorrect effect size estimate in the original meta-analysis. (n=4 effect sizes from two studies corrected)

- **Inappropriate inclusion of experiments or study conditions.** Some experiments or study conditions were included incorrectly. For example, in one instance, the original meta-analysis treated conference papers or theses and corresponding journal articles as separate entries, despite the fact that these papers reported on the same data. In other cases, study conditions were included that did not represent a test of infant-directed speech preference. We also excluded one study that included a highly atypical dependent measure of IDS preference - manipulating a physical toy - that differed substantially from all other included studies. (n=6 effect sizes from three studies removed)

- **Inappropriate exclusion of experiments or study conditions.** For some papers, the original meta-analysis reported only a subset of the experiments or study conditions in a given paper that represented a test of IDS preference. For example, in one instance, a paper included 11 separate conditions evaluating IDS preference across all experiments,

but only 6 of these conditions were included in the original MA. (n=10 effect sizes from three studies added)

- **Inaccuracies or inconsistencies in moderator variables.** We also encountered many instances in which moderator variables were coded incorrectly or inconsistently across studies. For example, papers using the same stimulus set were sometimes coded as differing on a stimulus dimension (such as whether the speaker had experience interacting with children).

To address these issues, we sought to revise the MA to match the information in the original papers. A group of six coders inspected each individual paper, documented each issue they identified in the original MA, and proposed a solution to the issue. Then, at least one other coder reviewed the issue and discussed the best solution with the first coder. Any issues that involved a substantial change to an effect size estimate (due to an incorrect effect size, or due to including or excluding a particular study condition) were discussed and agreed upon by the entire coding group. Whenever possible, we re-computed effect sizes from information reported in the source paper. When the paper included insufficient information to derive an effect size estimate, we used the effect size reported in the appendix of the (Dunst et al., 2012) meta-analysis, so long as the effect size was not clearly inconsistent with the results reported in the original paper. Incorrectly included effect sizes were removed and inappropriately excluded effect sizes were included in the updated meta-analysis. In total, the effect sizes reported for 8 of the 16 studies in the (Dunst et al., 2012) meta-analysis were revised, with 10 effect sizes included in the original meta-analysis altered (6 removed and 4 corrected) and 10 new effect sizes from the original studies added in the revised meta-analysis (Table 1). All moderator variables were updated to reflect the information reported in the original paper. For a full overview of all issues we encountered during the re-coding process and each corresponding change that was made to the original meta-analysis, see `https://docs.google.com/spreadsheets/d/e/2PACX-1vQaFJkLsVlZNhj8-L8FJ3rmEkfKg5KHZALCMLrp9ki7Fbd9n5xhGGvLGsKQKB296gL8QlFIMq3c-nF7/pubhtml`.

**Table 1:** *Overview over the main revisions to Dunst et al. (2012)*

| Study ID | Main Issue | Solution | Explanation | Original No. Effects | Updated No. Effects |
|---|---|---|---|---|---|
| Cooper1990 | none | none | No major issues | 2 | 2 |
| Cooper1994 | incorrect effect sizes and sample sizes | correct sample sizes/ design coding and recompute effect sizes based on means and SDs extracted from the figure | Most of the key issues relate to Experiment 3, which was previously incorrectly coded/ handled as a between-subjects condition and where sample sizes were incorrect (20 *total* infants, split between an IDS-first and ADS-first group, so n=10 per row), leading to some likely inaccurate effect size estimates. The sample size values have been corrected and means/ SDs were extracted from the figures in order to recompute the effect sizes. | 4 | 4 |
| Cooper1997 | none | none | No major issues | 3 | 3 |

| | | | | | |
|---|---|---|---|---|---|
| Fernald1985 | incorrect effect sizes in Dunst | recomputed effect sizes | Effect size reported by Dunst is d=0.65. However, after extracting data based on Figure 3, the effect size was recomputed as d=0.32 | 1 | 1 |
| Fernald1987 | incorrect effect sizes in Dunst | recomputed effect sizes | The effect sizes for all three experiments reported in Dunst are incompatible with the data reported in the paper. Original data was determined based on Ferrnald's dissertation (Figure 5) and effect sizes were recomputed in each case | 3 | 3 |
| Glenn1983 | atypical procedure and implausible effect sizes | removed experiment | Two main issues: (a) the procedure is highly unusual and is not a looking-time-based procedure (unlike every other study in the meta-analysis). Therefore, the study should be excluded based on its procedure type. (b) The effect sizes for this study are extremely large ($ds \approx 2.5$), and insufficient information is provided in the original paper to determine how they were computed by Dunst. Given these two major issues, we decided to exclude this study. | 2 | 0 |
| Kaplan1995a | none | none | Only minor issues not directly related to effect size (though effect size cannot be recomputed from data available in the paper) | 2 | 2 |
| Kaplan1995b | removed effects from trial 10 | removed data points from experiment | Audio (IDS/ ADS) was presented only on trial 9, hence only data from this trial is included. Three subsequent trials (trials 10, 11 and 12) are presented in silence. Dunst included the first of these (trial 10) - however, it is not clear why trials 11 and 12 were not also included. In our view, there are two defensible positions: include only trial 9 (only trial with audio stimulus), or all trials 9-12 (post-audio). We think the former option is the most straightforward (include only trial 9) | 4 | 2 |
| Pegg1989 | duplicate study | removed experiment | This paper is a conference proceedings paper using data that was eventually published in Pegg1992 (with a high, high likelihood). We therefore included only Pegg1992 and removed Pegg1989. | 2 | 0 |
| Pegg1992 | none | none | One key issue cannot be resolved (cannot determine non-significant effect sizes in Exp 1) | 2 | 2 |
| Schachner2011 | none | none | Only minor issues not directly related to effect size | 2 | 2 |
| Singh2002 | missing studies | add previously omitted studies testing IDS/ ADS preference | Experiments 1, 4, and 5 were previously not included in the meta-analysis, despite testing an IDS vs. ADS difference | 6 | 11 |
| Singh2009 | missing conditions | add previously omitted condition testing IDS/ ADS preference | Comparison of unfamiliar passages previously not included, despite also providing a test of IDS vs ADS preference | 2 | 3 |
| Trainor1996 | missing conditions | split data into six individual between-subjects conditions | Previous coding did not distinguish between different conditions of the experiment and appears to have omitted conditions with effects in the opposite direction, leading to an inflated effect size estimate; solution is to code each of the six conditions (each presenting a different IDS/ ADS stimulus set) as separate rows/ effect sizes | 2 | 6 |
| Werker1989 | none | none | One key issue cannot be resolved (cannot determine non-significant effect size in Exp 3) | 6 | 6 |

| | | | | | |
|---|---|---|---|---|---|
| Werker1994 | none | none | No large issues related to effect size. Effect sizes will be recomputed from Ms/ SDs derived from figures | 8 | 8 |

## 2.2. Augmenting the meta-analysis

After the meta-analysis data from Dunst et al. (2012) had been digitized for use on MetaLab (`https://langcog.github.io/metalab`), it was open to community-augmentation (i.e. members of the community could propose additional relevant papers). These papers were screened and added by a data manager. Additionally, experts from the field could suggest papers to add. In addition to ad-hoc additions, new studies were primarily added in two new literature search waves conducted in 2017 and 2019 in Google Scholar using a reverse-citation approach, identifying all publications citing two early studies of IDS (Fernald (1985); Cooper & Aslin (1990). Studies were screened for inclusion by trained community members. The final set of studies were coded by the same members of the authorship team who also conducted the revision of the Dunst et al. (2012) meta-analysis (Section 2.1) to ensure continuity in the coding process. All papers were coded by at least two team members. All coding discrepancies were documented and resolved through discussion among the entire coding team. Overall, this process resulted in a community-augmented meta-analysis comprising 30 studies contributing a total of 112 estimates.

# 3. Methodological details for ManyBabies 1

## 3.1. Sampling

Over the course of 14 months, labs were asked to test infants in up to four age groups (3-6, 6-9, 9-12, 12-15 months of age) and contribute at least 16 potentially eligible participants (before experimental exclusions, such as not enough data). Participants were tested across four continents (North America, Europe, Asia, Australia), and grew up learning one of 12 different languages, and therein four different varieties of English; of which two were classified as North-American (Canadian and US English) and 2 as non-North American English (Australian and British English). Since the stimuli were in North American English, only North American English learning participants were considered to be listening to native speech. All participants were monolingual; we are not including the data from the bilingual sample sister project (Byers-Heinlein et al., 2021).

## 3.2. Exclusion Criteria

Participant exclusion criteria included: (1) younger than 3 months or older than 15 months; (2) a known developmental delay; (3) premature birth (before 37 weeks); (3) experimenter error; (4) no usable trials (less than 1 trial per condition with at least 2 s total looking time to the screen). Trials were excluded (1) when the minimal looking criterion of 2 s was not met or the infant was inattentive/fussy during the trial; (2) due to technical errors; and (3) because of parental interference. Our dataset has all trial-level exclusions already applied

and thus follows the published report; for detailed exclusion statistics we refer to the paper reporting on the replication (The ManyBabies Consortium, 2020).

# 4. Moderators

## 4.1. Moderator Descriptions

The moderators included in our confirmatory analyses vary in their theoretical importance. We discuss them in the order reflecting the expected magnitude of impact they have on infant performance. Note that the MLR only varies in the first three of these moderators. We were only able to successfully fit models including these first three moderators, as reported in the main text.

**Age** Age is a key factor in developmental phenomena, which often emerge and change over time. Both the MLR and MA report a positive age effect, such that older infants show a larger preference. Theoretically, age could affect the measured preference for IDS in various directions. Younger infants might be expected to show increased preference for IDS due either to greater focus on broad acoustic characteristics of the speech or due to the greater importance of IDS pedagogically early in development. However, older infants become more mature language processors and accumulate language experience, which might allow them to more easily "tune in" to features of IDS. Older infants are also more cognitively and physically mature and might logically be expected to "perform" better in laboratory experiments more generally for reasons unrelated to the specific phenomenon under investigation (e.g., attentional control, comprehension of social expectations) (see e.g., Bergmann et al., 2018).

**Test language** While there are reasons to believe that some characteristics of IDS are universal, there is variation in their realization in different languages (Fernald et al., 1989; Cox et al., 2022). As a consequence, infants' preference for IDS may vary depending on whether or not stimuli were presented in infants' native language. Furthermore, even if IDS itself were universally specified, infants hearing speech stimuli in a non-native language may devote attentional resources differently than those hearing speech in their native language. Indeed, ManyBabies 1 reports that the IDS preference is stronger when the stimuli matched infants' native tongue.

**Experimental method** Method effects have been shown across tasks and ages, for example when pooling over 12 meta-analyses on early language acquisition (Bergmann et al., 2018). We thus expect an effect of method to be present in the aggregated data as well, among other factors because the original Dunst et al. (2012) MA was part of the pooled datasets for the just-cited meta-MA and because the current MLR found method effects. But whether these effects are consistent across datasets is unknown. We group method as follows: Headturn Preference Procedure (HPP), Central fixation (CF; including single-screen and eyetracking), Other (Forced Choice, FC; Conditioned Headturn, CHT). The categories are motivated by similarity in the tasks, i.e. either looking to vs away from a single screen with an unrelated visual display (CF), turning the head to the side towards flashing lights (HPP), or other tasks

which have only been used for a handful of estimates each (forced choice and conditioned headturn, both being used for four estimates, respectively). Note that the ManyBabies 1 study only included the first two of these three categories. ManyBabies 1 also included a third category, eye tracking. In the current analyses, we collapsed the methods termed "central fixation" and "eye tracking" in ManyBabies 1 into a single method category (CF), because we defined method in terms of the type of task procedure (as opposed to technical equipment used), and both central-fixation and eye-tracking experiments in ManyBabies 1 involved the same task procedure (i.e., a given trial worked exactly the same way from the perspective of an infant participant).

**Speech type** Conditions under which the stimuli were recorded were reported to influence effect size in the MA. This stimulus characteristic includes naturalistic (i.e. parents talking to their child), simulated (i.e. someone speaking *as if* talking to a child), and filtered or synthesized speech (i.e. manipulated recordings that sounded unnatural). As found in the original MA, we expect that the strength of the effect is highest for natural speech (used in the large-scale replication), followed by simulated speech, with filtered and synthesized speech showing smaller effects in turn. This effect might interact with age, but we did not include this interaction in the preregistered analyses because of power concerns and because we have insufficient grounds in the literature for strong predictions.

**Speaker familiarity** We expected a stronger effect for a highly familiar speaker, especially the infant's main caregiver (the infant's mother in the included studies). An advantage for maternal speech was reported by van Rooijen et al. (2019) and Barker & Newman (2004). These studies all compared the infant's own mother's voice to the voice of another infant's mother in the same study. In contrast to these experimental manipulations of speaker familiarity showing a benefit for the own mother's voice, however, Dunst et al. (2012) report a smaller effect size for the child's mother when comparing across studies. To further investigate this effect, we added whether the speaker was the child's own mother as a possible moderator.

**Mode of presentation** We tracked whether infants were presented with an unrelated visual stimulus or saw a video of a speaker, as this methodological variation might heighten infants' attention. This effect could either lead to an overall longer looking time across conditions (which would not be reflected in the effect size) or an increase in the difference between conditions (i.e. a larger effect).

**Dependent measure** Following the MA, we grouped the dependent variables into preference and affect. Studies either measured infants' looking time to a visual display (collapsing over the previous mentioned distinction between a related and unrelated visual stimulus) or infants' facial expression (e.g., smiling). All of these measures rely on behaviours that differ in the effort and conscious control the infant needs to exert (e.g., automatic smiling versus turning the head sideways and maintaining this position), and thus might impact the measured effect.

**Study goal** We coded whether infants' preference for IDS over ADS was the main research question of a paper since studies might also display the two types of speech stimuli to assess secondary phenomena, such as whether the presence or absence of IDS influences infants' preferences for specific speakers (Schachner & Hannon, 2011). While such studies contain the main comparison of interest, authors might add factors relevant to their research question, which in turn may lead to comparatively less controlled IDS and ADS stimuli. Since most stimuli are not available to us for direct comparison, we use the variable of whether IDS preference was a key research question as a proxy for stimulus quality.

## 4.2. Moderator Distributions

A key factor affecting our ability to fit moderator models is overall differences in the distribution of moderators between the MA and MLR, as well as differences in these distributions across multiple moderators (Tipton et al., 2019). To illustrate these differences, **Figure 1** shows the distribution of three key moderators (Infant Age, Method, and Test Language) for the MA and MLR. Notable patterns include: (1) only MA studies use artificial stimuli and all of these studies use methods that are neither HPP or CF; (2) MA studies using the HPP method exclusively test infants in their native language; and (3) MA studies have a wider age distribution, but this is especially true for studies using the CF method.

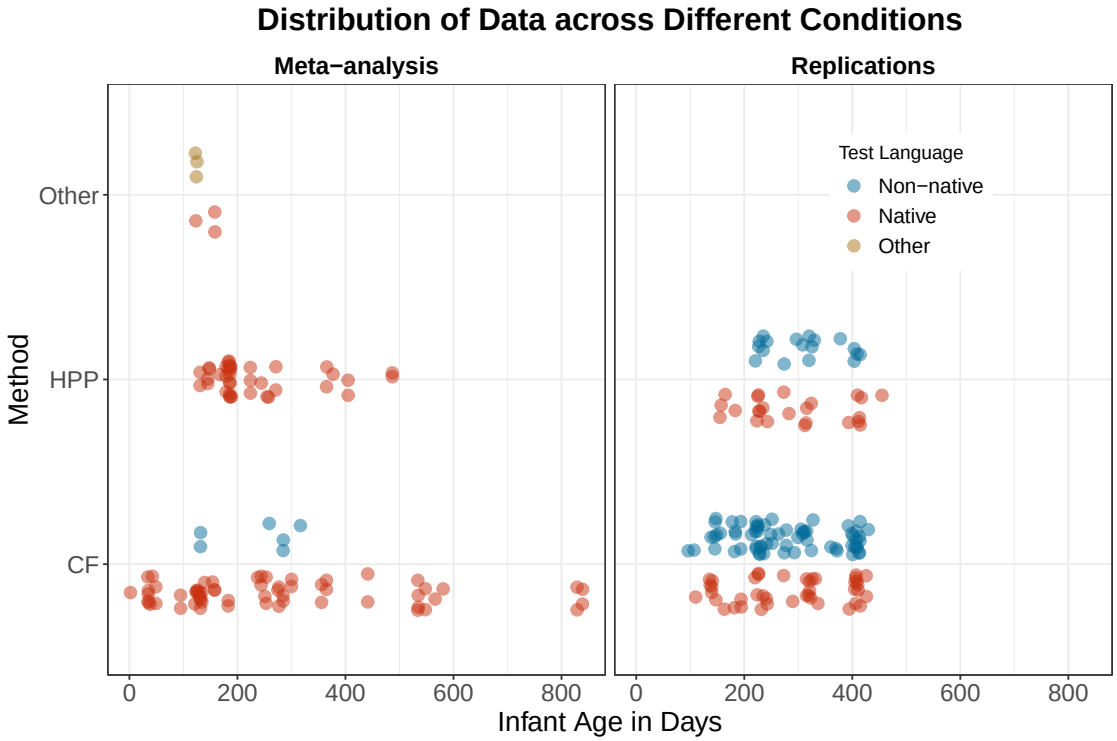


**Figure 1:** *Overview of distributions between moderators in the MA (left) and MLR (right). Studies depicted with a blue circle used non-native stimuli, studies depicted in orange used stimuli in infants' native language, and studies in brown used artificial stimuli.*

.

# 5. Supplementary Results

## 5.1. Results from the uncorrected meta-analysis

Based on our revision of the MA by Dunst et al. (2012) during coding of additional moderator variables, we noted several points of concern, such as the duplicate inclusion of data based on proceedings and journal papers (see section 2.1 in the Supplementary Materials for a full overview). Because our preregistration focused on the uncorrected meta-analysis, we report statistical results for transparency reasons; however, the results from this section should be treated with caution given the issues identified in the original Dunst et al. MA.

The average estimated effect size for only the studies from the original, uncorrected Dunst et al. MA was 0.67 (95% CI: [0.38, 0.95]; $p = 0.0003$), with considerable heterogeneity (estimated standard deviation of population effects $\widehat{\tau} = 0.51$). Nearly all of the population effects were estimated as positive (90% [85%, 97%]) and that nearly all were stronger than $SMD = 0.2$ (86% [80%, 93%]). Among only the MLR studies, the estimated average effect size was half as large (0.34 [0.27, 0.42]; $p<0.0001$) and with less estimated heterogeneity ($\widehat{\tau} = 0.11$). Despite the much smaller mean estimate in the MLR compared to the MA, we estimated that nearly all of the population effects were positive (100% [90%, 100%]) and that a large majority were stronger than $SMD = 0.2$ (89% [77%, 100%]), similar to the MA. This occurred because effects in the MLR were much more concentrated around their average than effects in the MA.

| **Statistical measure** | **Unadjusted model** | **Moderated model** |
|---|---|---|
| $\widehat{\mu}$ in MA | 0.64 [0.37, 0.91] | 0.58 [0.29, 0.88] |
| $\widehat{\mu}$ in MLR | 0.34 [0.27, 0.42] | 0.14 [-0.07, 0.34] |
| $\widehat{\mu}$ discrepancy | 0.29 [0.03, 0.56] | 0.45 [0.02, 0.88] |
| % effects $> 0$ in MA | 90 [85, 97] | 98 [91, 100] |
| % effects $> 0$ in MLR | 100 [90, 100] | 100 |
| % effects $> 0$ discrepancy | -10 [-15, -2] | -2 [-9, 0] |
| % effects $> 0.2$ in MA | 86 [80, 93] | 84 [81, 85] |
| % effects $> 0.2$ in MLR | 89 [77, 100] | 0 |
| % effects $> 0.2$ discrepancy | -3 [-19, 12] | 84 [-16, 87] |

**Table 2:** *$\widehat{\mu}$: Average effect size (SMD), as estimated in a meta-regression model containing both sources (original MA and MLR). % 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 $\widehat{\mu}$ measures (Hedges et al., 2010) and for differences in $\widehat{\mu}$ between sources and are bootstrapped for the percentage measures and for all cross-model comparisons (Mathur & VanderWeele, 2020a, 2021). Confidence intervals are omitted when they were not statistically estimable (i.e., for percentage estimates that were very close to 0% or 100%).*

In the **unadjusted model** that combined the two sources without any additional moderators, the estimated average effect sizes in the MA and in the MLR, respectively, were $SMD = 0.64$ (95% CI: [0.37, 0.91]) and 0.34 (95% CI: [0.27, 0.42]) (Table 2).[a] Thus, effect sizes in the MA were larger by on average 0.29 (95% CI: [0.03, 0.56]) units on the $SMD$ scale. There was considerable residual heterogeneity (estimated standard deviation of population effects $\widehat{\tau}_{\text{unadjusted}} = 0.32$).

The **moderated model** converged when we included three moderators besides source: infant age, test language, and method (Table 3). In the moderated model, 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 and test language) in the MA was, respectively, 0.58 [0.29, 0.88] and 0.14 [−0.07, 0.34]. Thus, effect sizes in the MA were larger by, on average, 0.45 [0.02, 0.88] $SMD$ units when controlling for these three moderators. This discrepancy was, if anything, larger than that seen in the unadjusted model, and the residual heterogeneity appeared essentially unchanged ($\widehat{\tau}_{\text{mod}} = 0.29$).

| Moderator | Est | CI | $p$-value |
|---|---|---|---|
| Intercept | 0.14 | [-0.07, 0.34] | 0.188 |
| Source: Meta-Analysis | 0.45 | [0.02, 0.88] | 0.041 |
| Infant Age (months) | 0.05 | [0.02, 0.07] | $< 0.001$ |
| Test Language: Non-native | -0.10 | [-0.21, 0.01] | 0.069 |
| Test Language: Other | -0.46 | [-2.40, 1.48] | 0.399 |
| Method: HPP | 0.11 | [-0.22, 0.43] | 0.506 |
| Method: Other | 0.60 | [-1.19, 2.38] | 0.300 |

**Table 3:** *Meta-regression estimates (original MA) 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 participant 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.*

## 5.2. Results from the revised Dunst et al. (2012) meta-analysis

After amending the original Dunst et al. (2012) MA in light of several concerns — but still retaining only the papers included in the Dunst et al. (2012) MA —, the overall effect size in the revised Dunst et al. (2012) MA was $SMD = 0.50$ [0.27, 0.73] and the MA continued to

[a] These estimates differed negligibly from those obtained by fitting separate models to the MA and MLR studies. Separate models are not exactly equivalent to meta-regression because, for example, separate models involve separate heterogeneity estimates whereas meta-regression has a single, average heterogeneity estimate. The heterogeneity estimate in turn slightly affects estimates' relative weights in the model.

contain substantial heterogeneity (estimated standard deviation of population effects $\widehat{\tau} = 0.32$). 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 ($\widehat{\tau} = 0.11$) compared to the MA. The meta-analytic effect size in the revised MA thus became more comparable to that of the MLR, with the difference between the two estimated as $SMD = 0.14$ [−0.08, 0.36] ($p = 0.21$). **Figure 2** provides an overview over the population effects for studies in both studies, while **Figure 3** shows the estimated densities for both the marginal and conditional population effects. Note in both plots the greater heterogeneity exhibited by the revised MA compared to MLR. As in the results in the main text, we also visualize the pooled point estimates for the subset of studies in the revised MA and in the MLR for each categorical candidate moderator (**Figure 4**).

Overall, we continued to observe a larger amount of heterogeneity in the MA and a numerical discrepancy between MA and MLR after revising the papers included in the Dunst et al. MA. We thus proceeded with the comparison of both datasets (Table 4). Despite the slightly smaller mean estimate in the MLR compared to the MA, we again estimated that nearly all of the population effects for the MLR and the MA were positive and that a large majority were stronger than $SMD = 0.2$.

| **Statistical measure** | **unadjusted model** | **Moderated model** |
|---|---|---|
| $\widehat{\mu}$ in MA | 0.48 [0.25, 0.71] | 0.49 [0.2, 0.77] |
| $\widehat{\mu}$ in MLR | 0.34 [0.27, 0.42] | 0.2 [0.04, 0.37] |
| $\widehat{\mu}$ discrepancy | 0.14 [-0.08, 0.36] | 0.28 [-0.06, 0.62] |
| % effects $> 0$ in MA | 87 [87, 92] | 98 [97, 99] |
| % effects $> 0$ in MLR | 100 [96, 100] | 100 |
| % effects $> 0$ discrepancy | -13 [-14, -8] | -2 [-3, -1] |
| % effects $> 0.2$ in MA | 78 [76, 88] | 93 [91, 96] |
| % effects $> 0.2$ in MLR | 89 [78, 100] | 0 |
| % effects $> 0.2$ discrepancy | -11 [-25, 12] | 93 [91, 96] |

**Table 4:** *$\widehat{\mu}$: Average effect size (SMD), as estimated in a meta-regression model containing both sources (revised MA and MLR). % 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 $\widehat{\mu}$ measures (Hedges et al., 2010) and for differences in $\widehat{\mu}$ between sources and are bootstrapped for the percentage measures and for all cross-model comparisons (Mathur & VanderWeele, 2020a, 2021). Confidence intervals are omitted when they were not statistically estimable (i.e., for percentage estimates that were very close to 0% or 100%).*

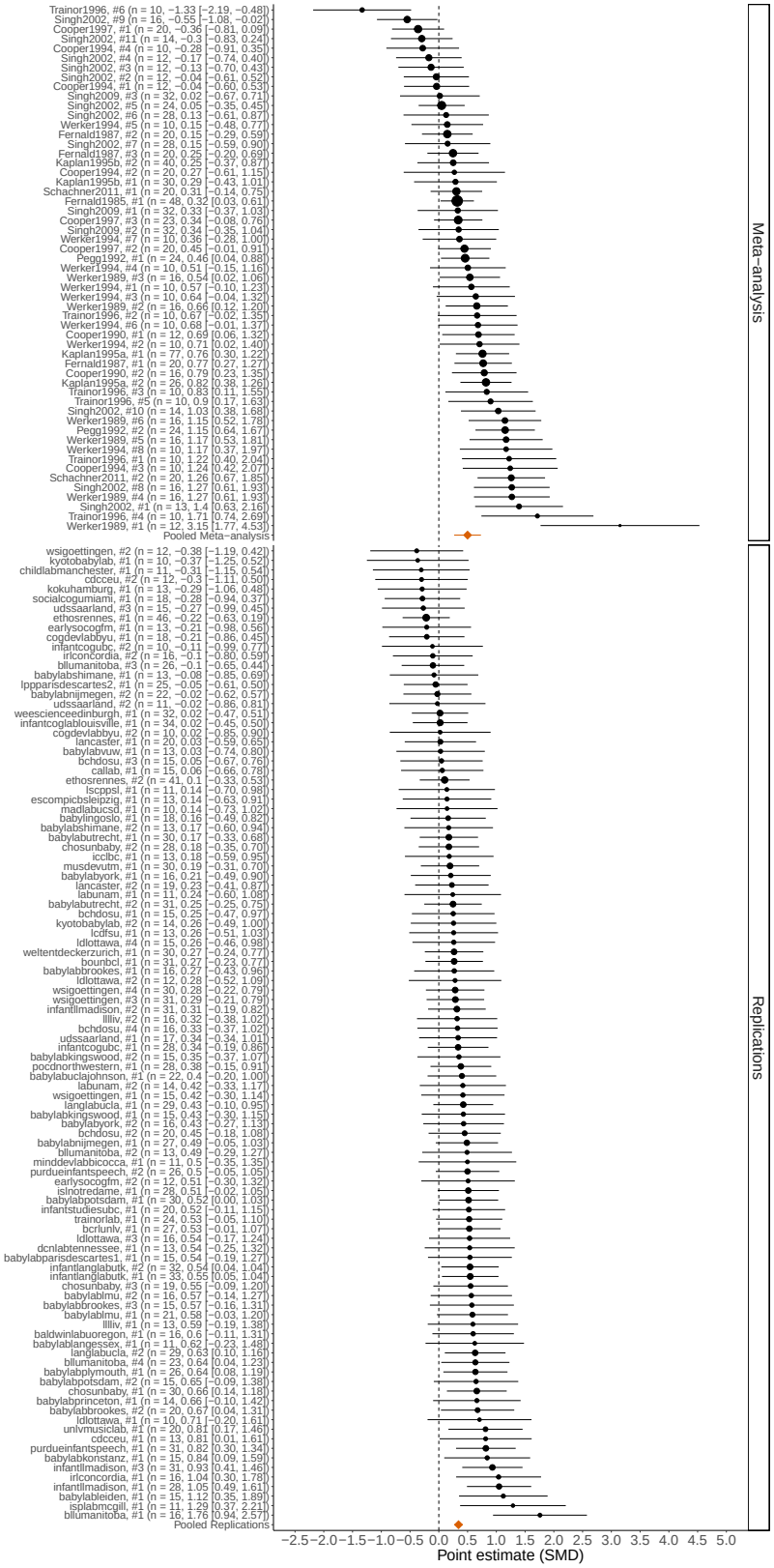


**Figure 2:** *Forest plot of studies' point estimates and 95% confidence intervals in the revised MA (top panel) and MLR (bottom panel). Orange diamond: pooled estimates within each source. Dashed vertical line: null.*

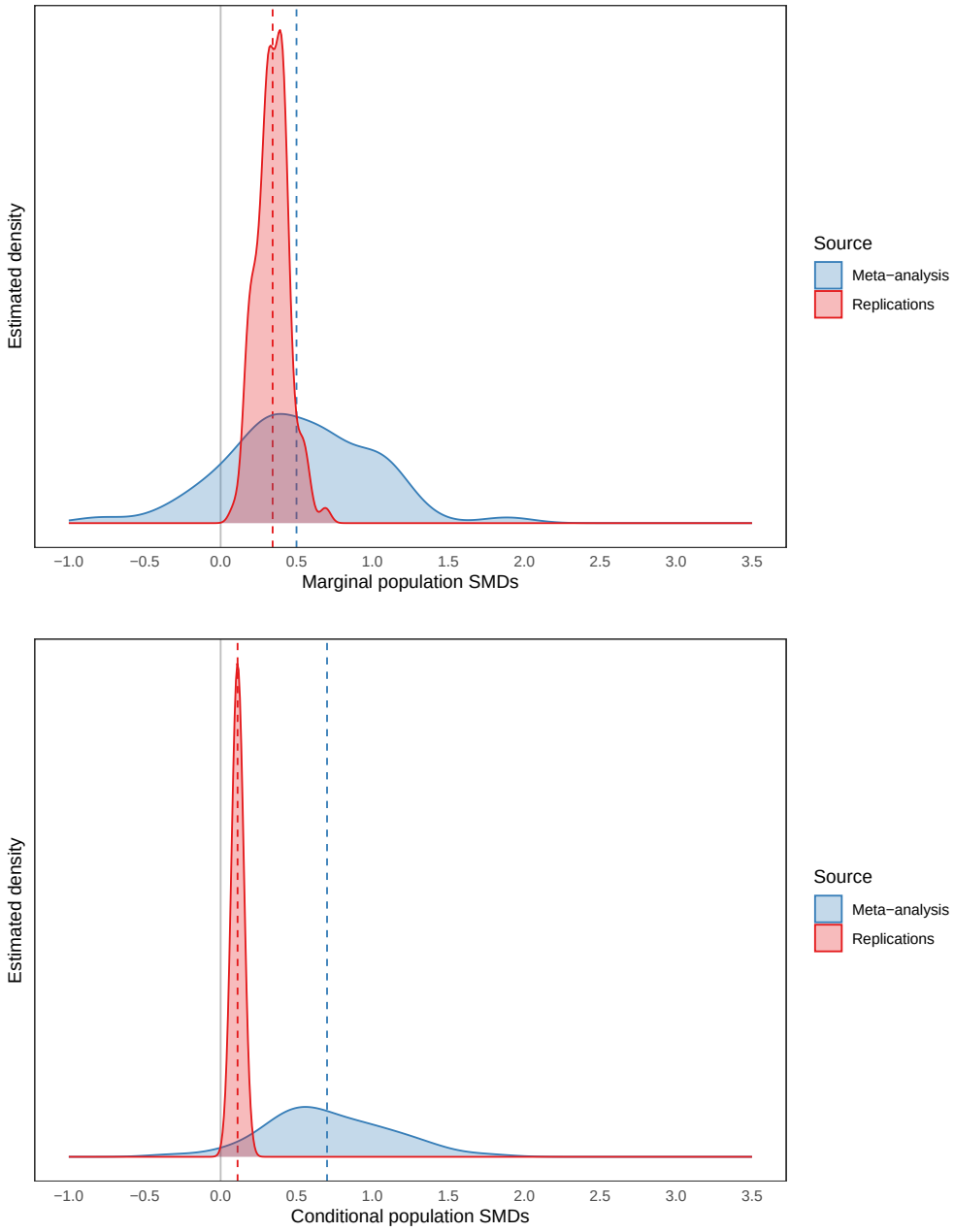


**Figure 3:** *Estimated densities of population effects in the revised MA (red) and in the MLR (gray). Top panel: Marginal population effects (i.e., not conditional on moderators). Bottom panel: Conditional population effects (i.e., conditional on the mean age and most common test language and method in the MA.) Vertical dashed lines: mean estimates from each source. Vertical gray line: null.*

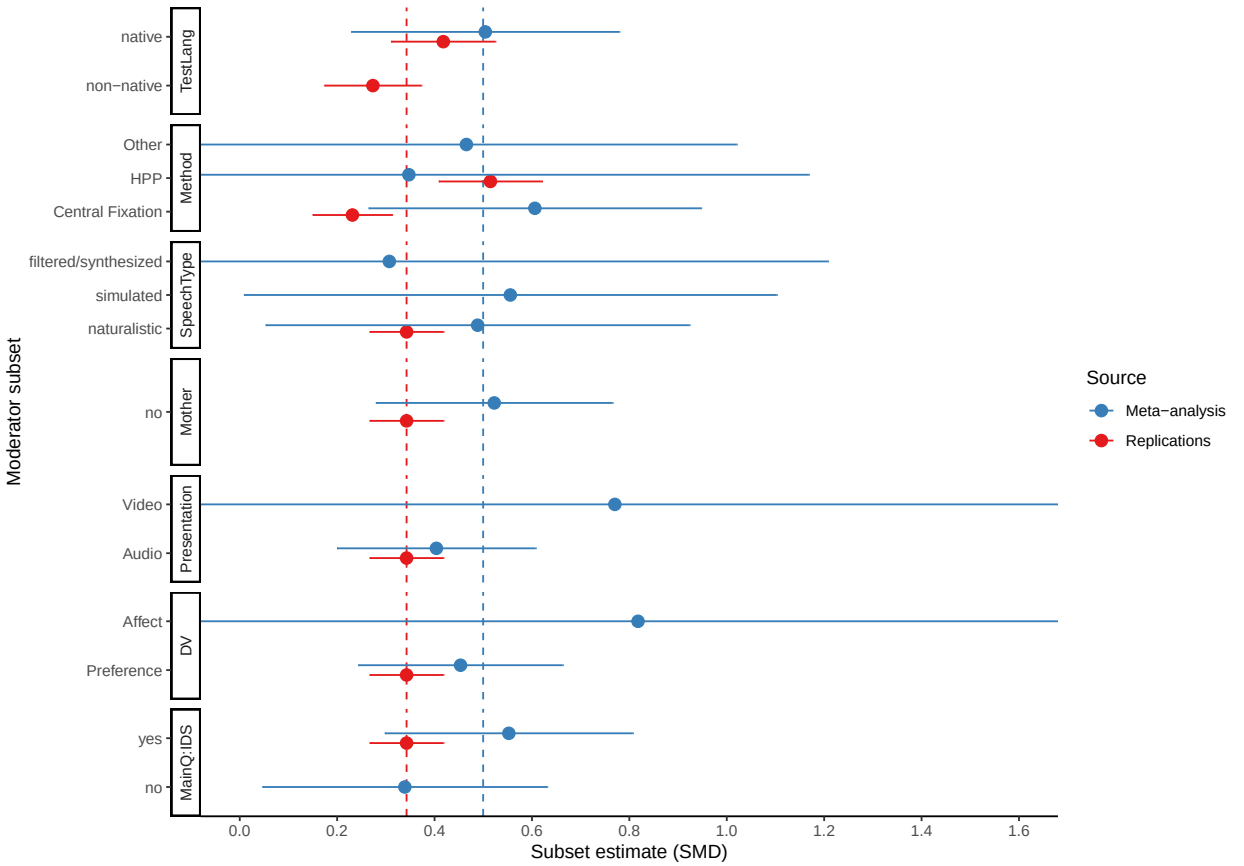


**Figure 4:** *Forest plot showing, for each categorical candidate moderator, the pooled point estimates for the subset of studies in the revised MA and in the MLR, respectively, with a given level of the moderator. Error bars are 95% confidence intervals. Error bars for many estimates are wide due to 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.*

### 5.2.1 Combined models

In the unadjusted model that combined the two sources without any additional moderators, effect sizes in the revised MA were larger by on average 0.14 (95% CI: [−0.08, 0.36]) units on the $SMD$ scale. There was considerable residual heterogeneity (estimated standard deviation of population effects $\widehat{\tau}_{\text{unadjusted}} = 0.27$).

The moderated model converged when we included the same three moderators as before (besides source): infant age, test language, and method. **Table 5** summarizes the estimates of the meta-regression for the remaining moderators, analogously to **Table 3**. The estimated average effect size in the revised MA and in the MLR when setting the moderators to their average values in the revised MA was, respectively, 0.49 [0.20, 0.77] and 0.20 [0.04, 0.37]. Thus, effect sizes in the MA were larger by, on average, 0.28 [−0.06, 0.62] $SMD$ units when controlling for these three moderators ($p = 0.09$). This discrepancy was, again, larger than that seen in the unadjusted model, and the residual heterogeneity appeared essentially unchanged ($\widehat{\tau}_{\text{mod}} = 0.28$).

| Moderator | Est | CI | $p$-value |
|---|---|---|---|
| Intercept | 0.2 | [0.04, 0.37] | 0.016 |
| Source: Meta-Analysis | 0.28 | [-0.06, 0.62] | 0.095 |
| Age (months) | 0.04 | [0.02, 0.07] | 0.001 |
| Test Language: Non-native | -0.10 | [-0.22, 0.02] | 0.105 |
| Test Language: Artificial | -0.15 | [-2.42, 2.12] | 0.558 |
| Method: HPP | 0.03 | [-0.21, 0.26] | 0.819 |
| Method: Other | 0.08 | [-1.75, 1.91] | 0.772 |

**Table 5:** *Meta-regression estimates (revised MA) 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 mean age, the estimate represents the increase in effect size associated with a 1-month increase in mean participant 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.*

#### 5.2.2 Publication bias

We next considered publication bias as a possible source of differences between the revised MA and MLR. The revised MA contained 25 affirmative and 30 nonaffirmative studies (vs 28 and 23, respectively in the original dataset). We began by implementing a correction for publication bias, estimating the level of affirmative selection from the MA itself. The average effect size in the MA after correction was $SMD = 0.41$ [0.18, 0.64]; $p = 5.0000 \cdot 10^{-4}$ (Vevea & Hedges, 1995), which was in this case indeed smaller than the uncorrected estimate of $SMD = 0.50$. 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 0.12 [−0.15, 0.39], which was in fact less than the estimate in the MLRs. Under "typical" publication bias in this field (favoring affirmative results by 4.7-fold), the MA average would decrease to 0.25 [−0.01, 0.50]. In both cases these estimates decreased sharply, were numerically smaller than MLR, and included zero in the 95% CI. We took this finding as suggesting that publication bias could therefore explain the discrepancy between the MA and MLR, which further motivated us to expand the MA to include papers and unpublished datasets beyond those included in the original Dunst et al. (2012) MA.

### 5.3. Additional analyses with the community-augmented meta-analysis

#### 5.3.1 Overview over included studies and effect sizes in the MA

**Table 6:** *Overview of the effect size estimates for each individual sample included in the community-augmented MA. Bracketed values are 95% confidence intervals.*

| Study | N | Mean Age (mos.) | Test Language | Method | Estimate | CI |
|---|---|---|---|---|---|---|
| Cooper & Aslin (1990) | 12 | 1.1 | Native | Central fixation | 0.69 | [0.06, 1.32] |
| Cooper & Aslin (1990) | 16 | 0.1 | Native | Central fixation | 0.79 | [0.23, 1.35] |
| Cooper & Aslin (1994) | 12 | 1.1 | Native | Central fixation | -0.04 | [-0.60, 0.53] |
| Cooper & Aslin (1994) | 20 | 1.1 | Native | Central fixation | 0.27 | [-0.61, 1.15] |
| Cooper & Aslin (1994) | 10 | 1.1 | Native | Central fixation | 1.24 | [0.42, 2.07] |
| Cooper & Aslin (1994) | 10 | 1.1 | Native | Central fixation | -0.28 | [-0.91, 0.35] |
| Cooper et al. (1997) | 20 | 1.3 | Native | Central fixation | -0.36 | [-0.81, 0.09] |
| Cooper et al. (1997) | 20 | 1.4 | Native | Central fixation | 0.45 | [-0.01, 0.91] |
| Cooper et al. (1997) | 23 | 4.1 | Native | Central fixation | 0.34 | [-0.08, 0.76] |
| Corbeil et al. (2016) | 20 | 8.5 | Non-Native | Central fixation | 0.10 | [-0.78, 0.98] |
| Droucker et al. (2013) | 22 | 6.0 | Native | Central fixation | 0.29 | [-0.14, 0.71] |
| Droucker et al. (2013) | 22 | 8.0 | Native | Central fixation | 0.29 | [-0.14, 0.72] |
| Droucker et al. (2013) | 22 | 12.0 | Native | Central fixation | 0.24 | [-0.18, 0.67] |
| Droucker et al. (2013) | 22 | 18.0 | Native | Central fixation | 0.52 | [0.08, 0.97] |
| Droucker et al. (2013) | 14 | 6.0 | Native | Central fixation | -0.01 | [-0.53, 0.52] |
| Droucker et al. (2013) | 14 | 8.0 | Native | Central fixation | 0.11 | [-0.42, 0.64] |
| Droucker et al. (2013) | 14 | 12.0 | Native | Central fixation | 0.29 | [-0.25, 0.82] |
| Droucker et al. (2013) | 14 | 18.0 | Native | Central fixation | 0.70 | [0.11, 1.28] |
| Fernald & Kuhl (1987) | 20 | 4.1 | Artificial | Other | 0.77 | [0.27, 1.27] |
| Fernald & Kuhl (1987) | 20 | 4.0 | Artificial | Other | 0.15 | [-0.29, 0.59] |
| Fernald & Kuhl (1987) | 20 | 4.1 | Artificial | Other | 0.25 | [-0.20, 0.69] |
| Fernald (1985) | 48 | 4.0 | Native | Other | 0.32 | [0.03, 0.61] |
| Hayashi et al. (2001) | 24 | 5.5 | Native | HPP | 0.77 | [0.32, 1.23] |
| Hayashi et al. (2001) | 31 | 8.5 | Native | HPP | 0.09 | [-0.26, 0.44] |
| Hayashi et al. (2001) | 34 | 12.4 | Native | HPP | 1.38 | [0.91, 1.85] |
| Inoue et al. (2011) | 17 | 8.0 | Native | HPP | 0.48 | [-0.03, 0.98] |
| Inoue et al. (2011) | 17 | 8.4 | Native | HPP | 0.12 | [-0.35, 0.60] |
| Kaplan et al. (1995a) | 77 | 4.1 | Native | Central fixation | 0.76 | [0.30, 1.22] |
| Kaplan et al. (1995a) | 26 | 4.1 | Native | Central fixation | 0.82 | [0.38, 1.26] |
| Kaplan et al. (1995b) | 30 | 4.1 | Native | Central fixation | 0.29 | [-0.43, 1.01] |
| Kaplan et al. (1995b) | 40 | 4.0 | Native | Central fixation | 0.25 | [-0.37, 0.87] |
| Kaplan et al. 2018 | 45 | 9.0 | Native | Central fixation | 0.31 | [0.02, 0.61] |
| Kaplan et al. 2018 | 21 | 8.2 | Native | Central fixation | -0.12 | [-0.55, 0.31] |
| Kim & Johnson (2014) | 42 | 5.2 | Native | Central fixation | 0.39 | [0.08, 0.71] |
| Kim & Johnson (2014) | 42 | 5.2 | Native | Central fixation | -0.12 | [-0.43, 0.18] |
| Kim & Johnson (2014) | 33 | 3.1 | Native | Central fixation | 0.43 | [0.08, 0.79] |
| Kim & Johnson (2014) | 33 | 3.1 | Native | Central fixation | 0.24 | [-0.11, 0.58] |
| McCartney (1997) | 24 | 4.3 | Native | Central fixation | -0.14 | [-0.54, 0.26] |
| McFayden et al. (2020) | 10 | 14.5 | Native | Central fixation | 0.63 | [-0.05, 1.31] |
| McFayden et al. (2020) | 10 | 14.5 | Native | Central fixation | 0.62 | [-0.06, 1.30] |
| Newman & Hussain (2006) | 30 | 4.3 | Native | HPP | 0.17 | [-0.19, 0.53] |
| Newman & Hussain (2006) | 30 | 4.3 | Native | HPP | 0.43 | [0.05, 0.80] |
| Newman & Hussain (2006) | 30 | 8.9 | Native | HPP | 0.06 | [-0.30, 0.41] |
| Newman & Hussain (2006) | 30 | 8.9 | Native | HPP | -0.16 | [-0.52, 0.20] |
| Newman & Hussain (2006) | 30 | 13.3 | Native | HPP | -0.04 | [-0.40, 0.32] |
| Newman & Hussain (2006) | 30 | 13.3 | Native | HPP | -0.23 | [-0.59, 0.13] |
| Newman (unpublished) | 24 | 4.9 | Native | HPP | -0.01 | [-0.41, 0.39] |
| Newman (unpublished) | 24 | 4.9 | Native | HPP | 0.11 | [-0.29, 0.52] |
| Newman (unpublished) | 15 | 4.8 | Native | HPP | -0.08 | [-0.59, 0.42] |
| Newman (unpublished) | 15 | 4.8 | Native | HPP | 0.19 | [-0.32, 0.70] |
| Ostroff (1998) | 20 | 10.4 | Non-Native | Central fixation | 0.55 | [0.08, 1.02] |
| Pegg et al. (1992) | 24 | 1.6 | Native | Central fixation | 0.46 | [0.04, 0.88] |
| Pegg et al. (1992) | 24 | 1.6 | Native | Central fixation | 1.15 | [0.64, 1.67] |
| Robertson et al. (2013) | 9 | 19.1 | Native | Central fixation | 0.95 | [0.16, 1.74] |
| Robertson et al. (2013) | 9 | 7.8 | Native | Central fixation | 0.38 | [-0.30, 1.05] |
| Robertson et al. (2013) | 9 | 18.6 | Native | Central fixation | 0.04 | [-0.61, 0.70] |
| Schachner & Hannon (2011) | 20 | 5.2 | Native | Other | 0.31 | [-0.14, 0.75] |
| Schachner & Hannon (2011) | 20 | 5.2 | Native | Other | 1.26 | [0.67, 1.85] |
| Segal & Newman (2015) | 36 | 12.0 | Native | HPP | 0.25 | [-0.08, 0.58] |
| Segal & Newman (2015) | 24 | 16.0 | Native | HPP | 0.35 | [-0.06, 0.76] |
| Segal & Newman (2015) | 36 | 12.0 | Native | HPP | 0.07 | [-0.26, 0.40] |
| Segal & Newman (2015) | 24 | 16.0 | Native | HPP | 0.01 | [-0.39, 0.41] |

| | | | | | | |
|---|---|---|---|---|---|---|
| Singh et al. (2002) | 13 | 6.1 | Native | HPP | 1.40 | [0.63, 2.16] |
| Singh et al. (2002) | 12 | 5.9 | Native | HPP | -0.04 | [-0.61, 0.52] |
| Singh et al. (2002) | 12 | 5.9 | Native | HPP | -0.13 | [-0.70, 0.43] |
| Singh et al. (2002) | 12 | 5.9 | Native | HPP | -0.17 | [-0.74, 0.40] |
| Singh et al. (2002) | 24 | 6.2 | Native | HPP | 0.05 | [-0.35, 0.45] |
| Singh et al. (2002) | 28 | 6.1 | Native | HPP | 0.13 | [-0.61, 0.87] |
| Singh et al. (2002) | 28 | 6.1 | Native | HPP | 0.15 | [-0.59, 0.90] |
| Singh et al. (2002) | 16 | 6.0 | Native | HPP | 1.27 | [0.61, 1.93] |
| Singh et al. (2002) | 16 | 6.0 | Native | HPP | -0.55 | [-1.08, -0.02] |
| Singh et al. (2002) | 14 | 6.1 | Native | HPP | 1.03 | [0.38, 1.68] |
| Singh et al. (2002) | 14 | 6.1 | Native | HPP | -0.30 | [-0.83, 0.24] |
| Singh et al. (2009) | 32 | 7.4 | Native | HPP | 0.33 | [-0.37, 1.03] |
| Singh et al. (2009) | 32 | 7.4 | Native | HPP | 0.34 | [-0.35, 1.04] |
| Singh et al. (2009) | 32 | 7.4 | Native | HPP | 0.02 | [-0.67, 0.71] |
| Trainor et al. (1996) | 10 | 6.1 | Native | HPP | 1.22 | [0.40, 2.04] |
| Trainor et al. (1996) | 10 | 6.1 | Native | HPP | 0.67 | [-0.02, 1.35] |
| Trainor et al. (1996) | 10 | 6.1 | Native | HPP | 0.83 | [0.11, 1.55] |
| Trainor et al. (1996) | 10 | 6.1 | Native | HPP | 1.71 | [0.74, 2.69] |
| Trainor et al. (1996) | 10 | 6.1 | Native | HPP | 0.90 | [0.17, 1.63] |
| Trainor et al. (1996) | 10 | 6.1 | Native | HPP | -1.33 | [-2.19, -0.48] |
| Wang et al. (2017) | 12 | 27.2 | Native | Central fixation | 0.87 | [0.21, 1.54] |
| Wang et al. (2017) | 22 | 11.7 | Native | Central fixation | 0.45 | [0.01, 0.89] |
| Wang et al. (2017) | 12 | 27.6 | Native | Central fixation | -0.02 | [-0.59, 0.54] |
| Wang et al. (2017) | 12 | 27.2 | Native | Central fixation | 0.66 | [0.04, 1.28] |
| Wang et al. (2017) | 22 | 11.7 | Native | Central fixation | 0.42 | [-0.01, 0.86] |
| Wang et al. (2017) | 12 | 27.6 | Native | Central fixation | 0.62 | [0.00, 1.24] |
| Wang et al. (2018) | 9 | 17.6 | Native | Central fixation | 0.09 | [-0.56, 0.75] |
| Wang et al. (2018) | 9 | 17.5 | Native | Central fixation | 0.26 | [-0.40, 0.93] |
| Wang et al. (2018) | 10 | 9.9 | Native | Central fixation | 0.84 | [0.12, 1.56] |
| Wang et al. (2018) | 14 | 9.1 | Native | Central fixation | 0.57 | [0.01, 1.14] |
| Wang et al. (2018) | 9 | 17.6 | Native | Central fixation | -0.06 | [-0.72, 0.59] |
| Wang et al. (2018) | 9 | 17.5 | Native | Central fixation | 0.34 | [-0.33, 1.01] |
| Wang et al. (2018) | 10 | 9.9 | Native | Central fixation | -0.83 | [-1.55, -0.11] |
| Wang et al. (2018) | 14 | 9.1 | Native | Central fixation | -0.16 | [-0.69, 0.37] |
| Ward & Cooper (1999) | 40 | 4.3 | Native | Central fixation | -0.68 | [-1.32, -0.04] |
| Ward & Cooper (1999) | 40 | 4.3 | Native | Central fixation | -0.63 | [-1.26, 0.01] |
| Werker & McLeod (1989) | 12 | 5.1 | Native | Central fixation | 3.15 | [1.77, 4.53] |
| Werker & McLeod (1989) | 16 | 4.4 | Native | Central fixation | 0.66 | [0.12, 1.20] |
| Werker & McLeod (1989) | 16 | 4.4 | Native | Central fixation | 0.54 | [0.02, 1.06] |
| Werker & McLeod (1989) | 16 | 8.3 | Native | Central fixation | 1.27 | [0.61, 1.93] |
| Werker & McLeod (1989) | 16 | 8.3 | Native | Central fixation | 1.17 | [0.53, 1.81] |
| Werker & McLeod (1989) | 16 | 4.6 | Native | Central fixation | 1.15 | [0.52, 1.78] |
| Werker et al. (1994) | 10 | 4.3 | Non-Native | Central fixation | 0.57 | [-0.10, 1.23] |
| Werker et al. (1994) | 10 | 4.3 | Non-Native | Central fixation | 0.71 | [0.02, 1.40] |
| Werker et al. (1994) | 10 | 9.4 | Non-Native | Central fixation | 0.64 | [-0.04, 1.32] |
| Werker et al. (1994) | 10 | 9.4 | Non-Native | Central fixation | 0.51 | [-0.15, 1.16] |
| Werker et al. (1994) | 10 | 4.3 | Native | Central fixation | 0.15 | [-0.48, 0.77] |
| Werker et al. (1994) | 10 | 4.3 | Native | Central fixation | 0.68 | [-0.01, 1.37] |
| Werker et al. (1994) | 10 | 9.4 | Native | Central fixation | 0.36 | [-0.28, 1.00] |
| Werker et al. (1994) | 10 | 9.4 | Native | Central fixation | 1.17 | [0.37, 1.97] |

### 5.3.2 Overview over included studies and effect sizes in the multi-lab replication

**Table 7:** *Overview of the effect size estimates for each individual sample included in the MLR. Bracketed values are 95% confidence intervals.*

| Lab ID | N | Mean Age (mos.) | Test Language | Method | Estimate | CI |
|---|---|---|---|---|---|---|
| babylabbrookes | 16 | 13.6 | Non-Native | Central fixation | 0.27 | [-0.43, 0.96] |
| babylabbrookes | 20 | 4.5 | Non-Native | Central fixation | 0.67 | [0.04, 1.31] |
| babylabbrookes | 15 | 7.6 | Non-Native | Central fixation | 0.57 | [-0.16, 1.31] |
| babylabkingswood | 15 | 13.4 | Non-Native | HPP | 0.43 | [-0.30, 1.15] |
| babylabkingswood | 15 | 7.3 | Non-Native | HPP | 0.35 | [-0.37, 1.07] |
| babylabkonstanz | 15 | 7.7 | Non-Native | HPP | 0.84 | [0.09, 1.59] |
| babylablangessex | 11 | 7.3 | Non-Native | Central fixation | 0.62 | [-0.23, 1.48] |

| | | | | | | |
|---|---|---|---|---|---|---|
| babylableiden | 15 | 10.5 | Non-Native | HPP | 1.12 | [0.35, 1.89] |
| babylablmu | 21 | 13.6 | Non-Native | Central fixation | 0.58 | [-0.03, 1.20] |
| babylablmu | 16 | 10.6 | Non-Native | Central fixation | 0.57 | [-0.14, 1.27] |
| babylabnijmegen | 27 | 7.7 | Non-Native | HPP | 0.49 | [-0.05, 1.03] |
| babylabnijmegen | 22 | 10.8 | Non-Native | HPP | -0.02 | [-0.62, 0.57] |
| babylabparisdescartes1 | 15 | 13.3 | Non-Native | HPP | 0.54 | [-0.19, 1.27] |
| babylabplymouth | 26 | 10.5 | Non-Native | HPP | 0.64 | [0.08, 1.19] |
| babylabpotsdam | 30 | 10.1 | Non-Native | HPP | 0.52 | [0.00, 1.03] |
| babylabpotsdam | 15 | 10.1 | Non-Native | Central fixation | 0.65 | [-0.09, 1.38] |
| babylabprinceton | 14 | 13.6 | Native | HPP | 0.66 | [-0.10, 1.42] |
| babylabshimane | 13 | 4.9 | Non-Native | Central fixation | -0.08 | [-0.85, 0.69] |
| babylabshimane | 13 | 7.6 | Non-Native | Central fixation | 0.17 | [-0.60, 0.94] |
| babylabuclajohnson | 22 | 13.3 | Native | Central fixation | 0.40 | [-0.20, 1.00] |
| babylabutrecht | 30 | 7.4 | Non-Native | HPP | 0.17 | [-0.33, 0.68] |
| babylabutrecht | 31 | 10.7 | Non-Native | HPP | 0.25 | [-0.25, 0.75] |
| babylabvuw | 13 | 7.5 | Non-Native | Central fixation | 0.03 | [-0.74, 0.80] |
| babylabyork | 16 | 7.2 | Non-Native | Central fixation | 0.21 | [-0.49, 0.90] |
| babylabyork | 16 | 10.4 | Non-Native | Central fixation | 0.43 | [-0.27, 1.13] |
| babylingoslo | 18 | 7.5 | Non-Native | Central fixation | 0.16 | [-0.49, 0.82] |
| baldwinlabuoregon | 16 | 10.5 | Native | Central fixation | 0.60 | [-0.11, 1.31] |
| bchdosu | 15 | 13.5 | Native | Central fixation | 0.25 | [-0.47, 0.97] |
| bchdosu | 20 | 4.6 | Native | Central fixation | 0.45 | [-0.18, 1.08] |
| bchdosu | 15 | 8.0 | Native | Central fixation | 0.05 | [-0.67, 0.76] |
| bchdosu | 16 | 10.6 | Native | Central fixation | 0.33 | [-0.37, 1.02] |
| bcrlunlv | 27 | 13.5 | Native | Central fixation | 0.53 | [-0.01, 1.07] |
| bllumanitoba | 16 | 13.6 | Native | HPP | 1.76 | [0.94, 2.57] |
| bllumanitoba | 13 | 5.4 | Native | HPP | 0.49 | [-0.29, 1.27] |
| bllumanitoba | 26 | 7.7 | Native | HPP | -0.10 | [-0.65, 0.44] |
| bllumanitoba | 23 | 10.3 | Native | HPP | 0.64 | [0.04, 1.23] |
| bounbcl | 31 | 13.4 | Non-Native | Central fixation | 0.27 | [-0.23, 0.77] |
| callab | 15 | 11.0 | Native | Central fixation | 0.06 | [-0.66, 0.78] |
| cdcceu | 13 | 13.1 | Non-Native | Central fixation | 0.81 | [0.01, 1.61] |
| cdcceu | 12 | 4.8 | Non-Native | Central fixation | -0.30 | [-1.11, 0.50] |
| childlabmanchester | 11 | 7.9 | Non-Native | Central fixation | -0.31 | [-1.15, 0.54] |
| chosunbaby | 30 | 13.3 | Non-Native | HPP | 0.66 | [0.14, 1.18] |
| chosunbaby | 28 | 7.5 | Non-Native | HPP | 0.18 | [-0.35, 0.70] |
| chosunbaby | 19 | 9.7 | Non-Native | HPP | 0.55 | [-0.09, 1.20] |
| cogdevlabbyu | 18 | 4.5 | Native | Central fixation | -0.21 | [-0.86, 0.45] |
| cogdevlabbyu | 10 | 6.4 | Native | Central fixation | 0.02 | [-0.85, 0.90] |
| dcnlabtennessee | 13 | 13.3 | Native | Central fixation | 0.54 | [-0.25, 1.32] |
| earlysocogfm | 13 | 7.4 | Native | Central fixation | -0.21 | [-0.98, 0.56] |
| earlysocogfm | 12 | 10.4 | Native | Central fixation | 0.51 | [-0.30, 1.32] |
| escompicbsleipzig | 13 | 5.1 | Non-Native | Central fixation | 0.14 | [-0.63, 0.91] |
| ethosrennes | 46 | 4.8 | Non-Native | Central fixation | -0.22 | [-0.63, 0.19] |
| ethosrennes | 41 | 7.7 | Non-Native | Central fixation | 0.10 | [-0.33, 0.53] |
| icclbc | 13 | 7.6 | Native | Central fixation | 0.18 | [-0.59, 0.95] |
| infantcoglablouisville | 34 | 10.7 | Native | Central fixation | 0.02 | [-0.45, 0.50] |
| infantcogubc | 28 | 4.6 | Native | Central fixation | 0.34 | [-0.19, 0.86] |
| infantcogubc | 10 | 7.4 | Native | Central fixation | -0.11 | [-0.99, 0.77] |
| infantlanglabutk | 33 | 13.7 | Native | HPP | 0.55 | [0.05, 1.04] |
| infantlanglabutk | 32 | 7.4 | Native | HPP | 0.54 | [0.04, 1.04] |
| infantllmadison | 28 | 13.5 | Native | HPP | 1.05 | [0.49, 1.61] |
| infantllmadison | 31 | 7.3 | Native | HPP | 0.31 | [-0.19, 0.82] |
| infantllmadison | 31 | 10.4 | Native | HPP | 0.93 | [0.41, 1.46] |
| infantstudiesubc | 20 | 7.5 | Native | HPP | 0.52 | [-0.11, 1.15] |
| irlconcordia | 16 | 13.0 | Native | Central fixation | 1.04 | [0.30, 1.78] |
| irlconcordia | 16 | 7.2 | Native | Central fixation | -0.10 | [-0.80, 0.59] |
| islnotredame | 28 | 13.5 | Native | HPP | 0.51 | [-0.02, 1.05] |
| isplabmcgill | 11 | 13.6 | Non-Native | HPP | 1.29 | [0.37, 2.21] |
| kokuhamburg | 13 | 7.4 | Non-Native | Central fixation | -0.29 | [-1.06, 0.48] |
| kyotobabylab | 10 | 7.3 | Non-Native | Central fixation | -0.37 | [-1.25, 0.52] |
| kyotobabylab | 14 | 9.8 | Non-Native | Central fixation | 0.26 | [-0.49, 1.00] |
| labunam | 11 | 13.3 | Non-Native | Central fixation | 0.24 | [-0.60, 1.08] |
| labunam | 14 | 7.3 | Non-Native | Central fixation | 0.42 | [-0.33, 1.17] |
| lancaster | 20 | 13.4 | Non-Native | Central fixation | 0.03 | [-0.59, 0.65] |
| lancaster | 19 | 7.4 | Non-Native | Central fixation | 0.23 | [-0.41, 0.87] |
| langlabucla | 29 | 5.1 | Native | HPP | 0.43 | [-0.10, 0.95] |
| langlabucla | 29 | 10.7 | Native | HPP | 0.63 | [0.10, 1.16] |

| | | | | | | |
|---|---|---|---|---|---|---|
| lcdfsu | 13 | 10.5 | Native | Central fixation | 0.26 | [-0.51, 1.03] |
| ldlottawa | 10 | 13.6 | Native | Central fixation | 0.71 | [-0.20, 1.61] |
| ldlottawa | 12 | 4.8 | Native | Central fixation | 0.28 | [-0.52, 1.09] |
| ldlottawa | 16 | 7.9 | Native | Central fixation | 0.54 | [-0.17, 1.24] |
| ldlottawa | 15 | 10.4 | Native | Central fixation | 0.26 | [-0.46, 0.98] |
| lllliv | 13 | 13.1 | Non-Native | Central fixation | 0.59 | [-0.19, 1.38] |
| lllliv | 16 | 7.4 | Non-Native | Central fixation | 0.32 | [-0.38, 1.02] |
| lppparisdescartes2 | 25 | 7.9 | Non-Native | HPP | -0.05 | [-0.61, 0.50] |
| lscppsl | 11 | 13.2 | Non-Native | Central fixation | 0.14 | [-0.70, 0.98] |
| madlabucsd | 10 | 7.7 | Native | Central fixation | 0.14 | [-0.73, 1.02] |
| minddevlabbicocca | 11 | 4.8 | Non-Native | Central fixation | 0.50 | [-0.35, 1.35] |
| musdevutm | 30 | 7.4 | Native | HPP | 0.19 | [-0.31, 0.70] |
| pocdnorthwestern | 28 | 13.4 | Native | Central fixation | 0.38 | [-0.15, 0.91] |
| purdueinfantspeech | 31 | 12.9 | Native | HPP | 0.82 | [0.30, 1.34] |
| purdueinfantspeech | 26 | 10.3 | Native | HPP | 0.50 | [-0.05, 1.05] |
| socialcogumiami | 18 | 4.5 | Native | Central fixation | -0.28 | [-0.94, 0.37] |
| trainorlab | 24 | 8.0 | Native | HPP | 0.53 | [-0.05, 1.10] |
| udssaarland | 17 | 13.4 | Non-Native | Central fixation | 0.34 | [-0.34, 1.01] |
| udssaarland | 11 | 7.8 | Non-Native | Central fixation | -0.02 | [-0.86, 0.81] |
| udssaarland | 15 | 10.4 | Non-Native | Central fixation | -0.27 | [-0.99, 0.45] |
| unlvmusiclab | 20 | 4.6 | Native | Central fixation | 0.81 | [0.17, 1.46] |
| weescienceedinburgh | 32 | 7.0 | Non-Native | Central fixation | 0.02 | [-0.47, 0.51] |
| weltentdeckerzurich | 30 | 13.6 | Non-Native | Central fixation | 0.27 | [-0.24, 0.77] |
| wsigoettingen | 15 | 13.5 | Non-Native | Central fixation | 0.42 | [-0.30, 1.14] |
| wsigoettingen | 12 | 4.8 | Non-Native | Central fixation | -0.38 | [-1.19, 0.42] |
| wsigoettingen | 31 | 7.3 | Non-Native | Central fixation | 0.29 | [-0.21, 0.79] |
| wsigoettingen | 30 | 10.2 | Non-Native | Central fixation | 0.28 | [-0.22, 0.79] |

### 5.3.3 Estimated densities of population effects in the MA and MLR

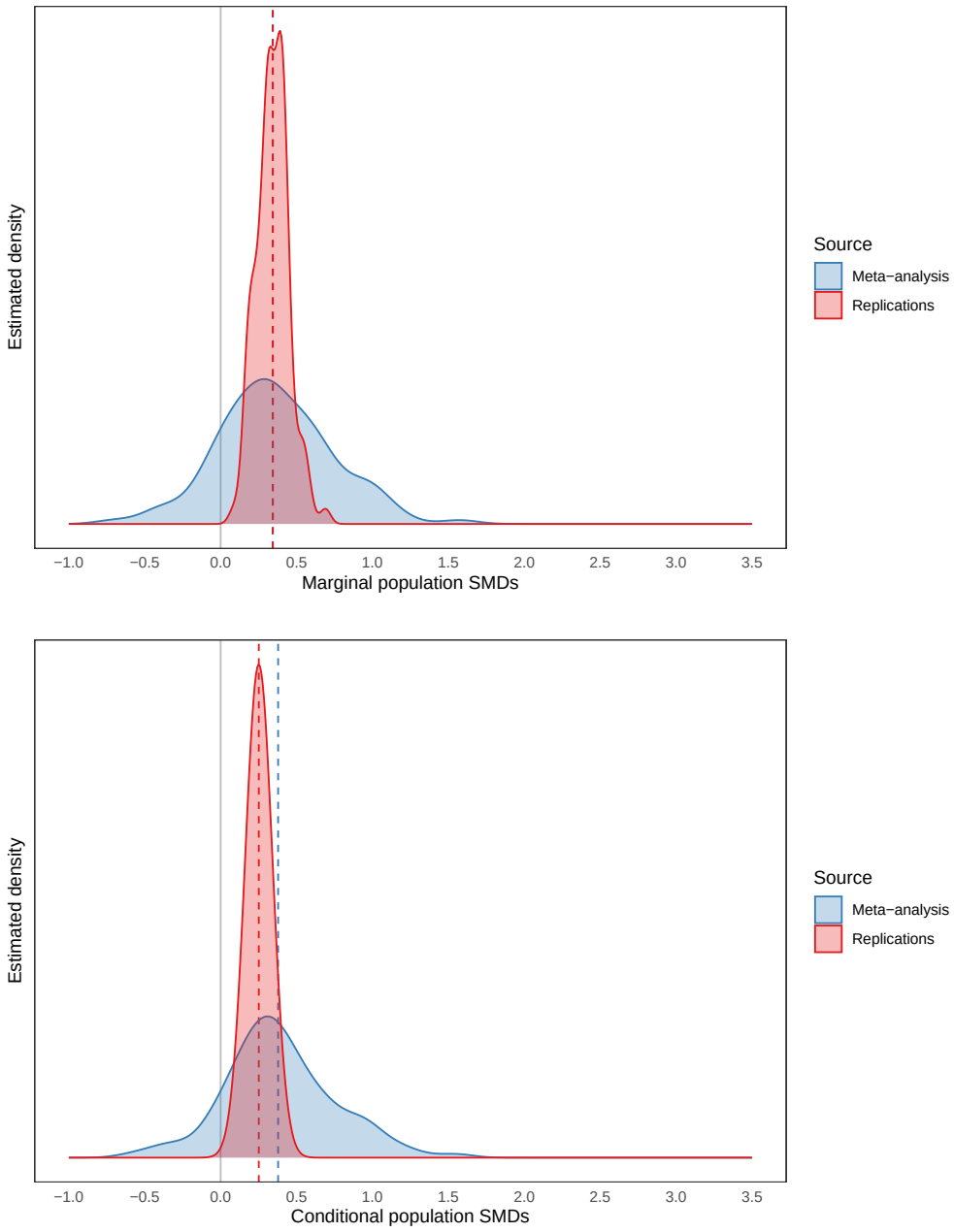


**Figure 5:** *Estimated densities of population effects in the community-augmented MA (red) and in the MLR (gray). Top panel: Unadjusted population effects (i.e., not conditional on moderators). Bottom panel: Conditional population effects (i.e., conditional on the mean infant age and most common test language and method in the MA.) Vertical dashed lines: mean estimates from each source. Note that in the top panel, the two estimates almost completely overlap. Vertical gray line: null.*

### 5.3.4 Sensitivity analysis: Within-subjects experiment design

A small number of studies in the MA used between-subjects rather than within-subjects designs. We anticipated that between-subjects designs may differ systematically from within-subjects designs because larger unintended variation between conditions in between-subjects designs is not necessarily countered by increasing the sample size in infant research, since testing is costly (Bergmann et al., 2018). We therefore repeated the analyses in Section 3.1 of the main manuscript after excluding from the MA the 12 estimates from between-subjects designs; doing so slightly increased the average effect size in the MA from 0.35 [0.22, 0.47] to 0.37 [0.24, 0.50]. The meta-regression results for both the unadjusted and moderated models corroborated the main results, suggesting that the use of between-subjects designs in some of the MA studies did not explain the discrepancies.

### 5.3.5 Additional analyses supporting publication bias methods

**Visual diagnostics for assumptions.** To assess for violations of the assumption that publication bias operates in favor of affirmative results (i.e., those with $p < 0.05$ and point estimates in the desired direction), we calculated and plotted one-tailed $p$-values from the meta-analysis studies (Figure 6). The much larger mass of one-tailed $p$-values below 0.025 (37.00% of all $p$-values) versus those above 0.975 (4.0% of $p$-values) suggested that any selection, if present, indeed was one-directional rather than two-directional.

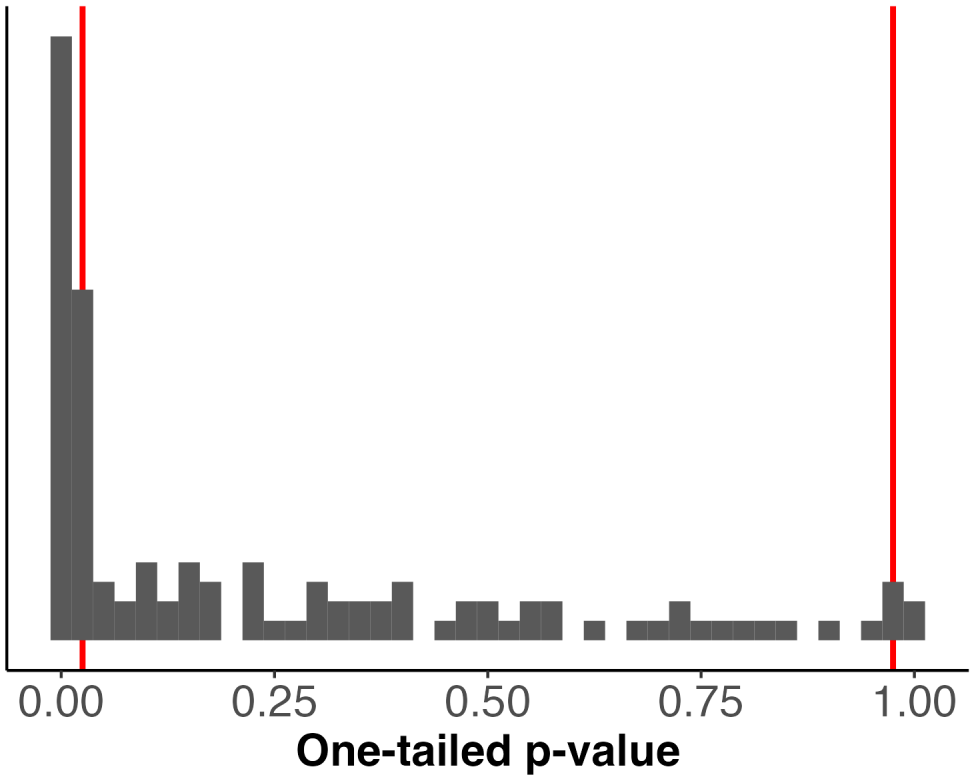


**Figure 6:** *One-tailed p-values from all studies in the meta-analysis. Red lines indicate the 0.025 and 0.975 thresholds, i.e., the thresholds at which the corresponding two-tailed p-value would be* $< 0.05$ *and in the desired direction and at which the two-tailed p-value would be* $< 0.05$ *but in the unanticipated direction.*

**Paper-reported significance.** Additionally, our main analyses defined statistical "significance" and affirmative status based on whether our calculated $p$-value was less than 0.05. Our calculated $p$-value sometimes differed from the $p$-value reported in the papers, likely due to differences in the statistical strategy used (e.g., the original study might have computed p-values using a one-tailed test) or (in some instances) due to a potential error in the original study. Of $p$-values that we calculated to be less than 0.05, authors reported 97% to be "significant"; and of $p$-values that we calculated to be greater than 0.05 (i.e., non-significant), authors in the original studies reported 29% to be "significant". We repeated the sensitivity analyses under the alternative assumption that publication bias favors results that studies' authors reported to be significant and positive (Mathur & VanderWeele, 2020b). These analyses yielded generally similar results, except that the estimate under hypothetical worst-case publication bias (0.04 [−0.11, 0.18]) was somewhat smaller than in main analyses (0.09 [−0.01, 0.18]).

### 5.3.6 Exploratory sensitivity analysis: Applying more stringent participant inclusion criteria in the multi-lab replication

Exploratory analyses in the MLR (The ManyBabies Consortium, 2020, Table 6) revealed that the observed effect size depends on the inclusion criterion applied to the data. In addition, a follow-up project assessing test-retest reliability found reliable effects across test sessions only with more stringent criteria (Schreiner et al., 2022). In infant studies, it is common to only include participants who contribute data in 80% or even 100% of the presented trials. Based on the few papers where inclusion criteria were reported in the MAs, we can infer that studies on IDS preference follow this general pattern. In contrast, the MLR used a criterion of a minimum 12.5% of the trials in their main analyses (1 trial per condition over 16 trials). This very loose criterion was used because the MLR included planned analyses regarding data loss. We opted for 75% of the data (i.e. 7 of 8 trials per condition) as a more stringent exclusion criterion, which is closer to literature standards – when reported – and which substantially reduced the amount of data analysed.

Applying the more stringent inclusion criterion left 54 estimates and 952 participants in the MLR. The estimated average effect size in the MLR increased somewhat to $SMD$ =0.42 [0.31, 0.53]. This average effect size was in fact numerically larger than that in the community-augmented MA by $SMD$ =−0.09 [−0.25, 0.07] ($p = 0.25$), but the confidence interval was wide. Other patterns in the main analyses are largely preserved.

### 5.3.7 Exploratory sensitivity analysis: Restricting the age range when investigating interactions between the MA and MLR

There was a wider range of infant ages among the studies included in the community-augmented MA (see Figure 4A in the main manuscript). We therefore also investigated the extent to which the results for the model investigating interactions between source and moderator variables held when only including studies within the same age range as the MLR (i.e., only including studies with an average participant age ranging between 3 and 15 months). The MA included $n$=86 studies after restricting the age range. We fit the same interaction model as in section 3.2 of the main manuscript, including the two-way interactions between source (MA vs. MLR) and each of the three key moderators (infant age, test language, and method). Overall, the results were similar to the results using the full dataset (cf., Table 8 and Figure 7). There was a significant interaction between source and method ($b = -0.36$ [−0.67, −0.06]; $p$ =0.02). We also found a marginal, non-significant interaction between source and infant age ($b = -0.04$ [−0.09, 0.00]; $p$ =0.06); however, the magnitude of the coefficient estimate was virtually identical to the estimate in the model including the full MA dataset. As in the model with the full dataset, there was no interaction between source and test language ($b = -0.19$ [−0.58, 0.20]; $p$ =0.27).

| Moderator | Est | CI | $p$-value |
|---|---|---|---|
| Intercept | 0.34 | [0.24, 0.44] | <0.0001 |
| Source (centered) | 0.03 | [-0.16, 0.23] | 0.698 |
| Age (months; centered) | 0.02 | [-0.00, 0.05] | 0.055 |
| Test Language (Native vs. Other) | 0 | [-0.20, 0.19] | 0.987 |
| Method (HPP vs. Other) | 0.05 | [-0.10, 0.21] | 0.486 |
| Source * Age | -0.04 | [-0.09, 0.00] | 0.061 |
| Source * Test Language | -0.19 | [-0.58, 0.20] | 0.265 |
| Source * Method | -0.36 | [-0.67, -0.06] | 0.021 |

**Table 8:** *Meta-regression estimates of the moderator interaction model when restricting the age range of the MA. 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 a binary variables and centered. Bracketed values are 95% confidence intervals.*

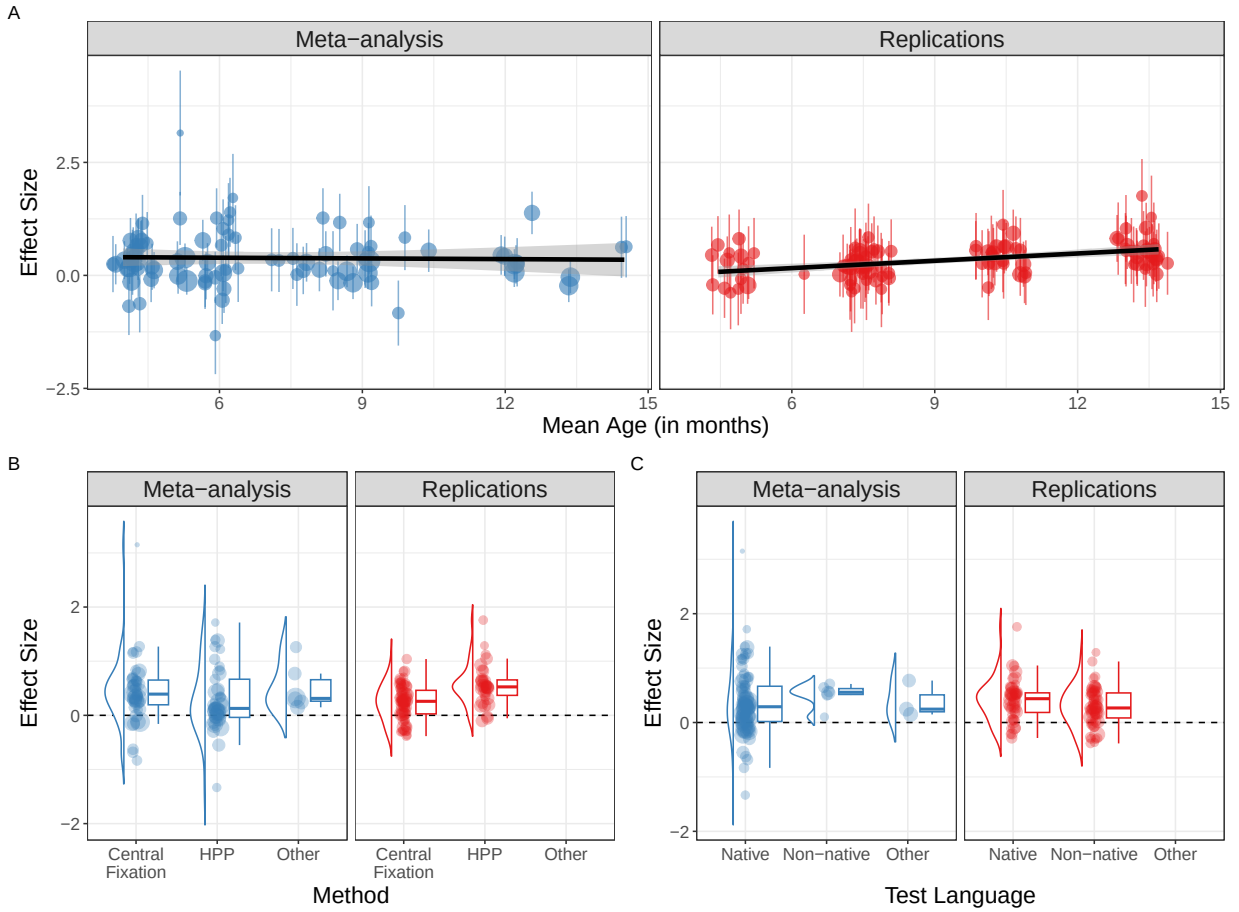


**Figure 7:** *Overview of the distribution of effect sizes in the meta-analytic dataset and the MLR restricted to studies with average ages between 3 and 15 months for three key moderators: age (A), method (B), and test language (C). In (A), the black line represents a linear fit through the effect sizes for each source and error bars for individual estimates are 95% confidence intervals.*

# 6. Discussion of discrepancies between results across MA datasets

In this section, we outline and contextualize the differences in results across the original, revised and community-augmented MA datasets (cf. Section 5.1 and 5.2 in Supplementary Materials to see the full results). Specifically, we clarify how the goals of the study and our interpretations changed with each update to the dataset and explain why we believe that the community-augmented MA provides the most comprehensive meta-analytic estimate currently available.

When we first planned to compare the results from the MA to the MLR, we were faced with an intriguing puzzle: the original Dunst meta-analysis found an effect size almost twice as large ($d = 0.67$) as the effect size in the ManyBabies MLR ($d = 0.35$). On the surface, this discrepancy was also consistent with other findings suggesting that MAs tend to inflate effect sizes relative to MLRs (Kvarven et al., 2020). What could explain such a large discrepancy? One likely explanation was publication bias: MAs could be inflated in part because non-significant studies are more likely to be file-drawered than significant studies, artificially raising the estimates in the pool of published effects. However, past analyses suggested that publication bias could not fully explain the discrepancies observed between MAs and MLRs (Lewis et al., 2022), leading us to consider an alternative possibility: perhaps differences between MLRs and MAs could be explained due to systematic variation in key moderators. Consistent with this idea, the ManyBabies 1 project found that the magnitude of effect sizes varied across several key design (e.g., experiment method) and participant characteristics (e.g., infant age) within the MLR. We reasoned that accounting for the effect of moderators might reduce differences in the estimates of IDS preference between the MA and MLR. When we evaluated this possibility (cf., Section 5.1.), however, we found that in fact the opposite was true: accounting for the effect of key moderators substantially increased the discrepancy in the estimates for the MA vs. the MLR, from 0.29 to 0.45 (Table 2). Publication bias on its own also could not fully account for the discrepancy between the MA and the MLR. Initially, therefore, we were faced with an even more puzzling result than the unadjusted difference in estimates between the MA and the MLR — not only could we not account for this dramatic difference by testing what appeared to be the most promising explanations, our attempts to account for sources of variation between the MA and MLR in fact made the difference more pronounced.

Part of the solution to this puzzle appears to be simply error in the original MA. In the course of re-coding moderator variables for the original MA, we discovered a series of decisions and errors that required revision. These issues ranged from including duplicated effect sizes, incompatibility between the effect sizes reported in the MA and data reported in the paper, and the omission of several experiments with non-significant findings from papers otherwise included in the meta-analysis (cf. Section 2 in Supplementary Materials). Revising the meta-analysis in light of these issues substantially lowered the overall effect size estimate (to 0.50). Using the revised meta-analysis, we still found that accounting for moderators did not reduce discrepancies between the MA and MLR (though we also no longer found that accounting for moderators significantly increased discrepancies). On the other hand, we found that — across several methods — accounting for publication bias could in fact

explain the remaining discrepancy between the MA and the MLR. Under the assumption of "typical" publication bias in the field, the MA effect size estimate was numerically smaller than the MLR. Given the documented inconsistencies in the original Dunst meta-analysis, we are thus inclined to treat the puzzles raised by the comparison of the MLR to the original Dunst meta-analysis as entirely an artifact of errors in the original MA. Once these errors were revised, the meta-analytic estimate moved substantially closer to the MLR estimate and the puzzling increase in MA-MLR differences when accounting for moderator variables was no longer as marked.

While recoding and revising the Dunst et al. meta-analysis, the metalab community also worked on substantially augmenting the meta-analysis, leading to a dataset that included almost twice the number of studies in the revised meta-analysis. Comparing the MLR to this community-augmented MA clarified the picture dramatically: the updated meta-analytic estimate and the MLR estimate were almost completely aligned. By combining data from both a large-scale MA and MLR, we could therefore provide a more comprehensive picture of the generalizability of IDS preference, as reported in the main manuscript. At the same time, comparing these sources raised a new puzzle: unlike in the previous analyses, we now observed significant interactions between two key moderators (infant age and experiment method) and the data source. Both the Dunst et al. meta-analysis and the MLR originally reported a similar positive relationship between age and effect size magnitude, such that IDS preference increased with age in models combining the two data sources. In the community-augmented MA, we no longer observed an age effect, and any effect of method went, if anything, in the opposite direction from the increased effect size for the HPP method observed in the MLR.

Overall, we believe that the community-augmented MA provides the most accurate, comprehensive meta-analytic estimate currently available. The fact that its estimate converges with the effect size observed in the MLR increases confidence in the robustness and precision of our current "best estimate" of IDS preference of $d \approx 0.35$. At the same time, the differences in the moderating effects observed within each data source highlight the limitations of both the MA and the MLR to fully address questions of key theoretical and methodological interest, such as the developmental trajectory of IDS preference and its dependence on different methods for eliciting preference. We hope that these remaining puzzles can motivate future work in the field and catalyze improvements in how we approach knowledge building from both MAs and MLRs. Here we highlight the value of manipulating key variables of interest to discover how effects, such as IDS preference, vary across a broad terrain of theoretically relevant dimensions.

## References

Barker, B. A., & Newman, R. S. (2004). Listen to your mother! the role of talker familiarity in infant streaming. *Cognition*, *94*(2), B45–B53.

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.

Byers-Heinlein, K., Tsui, A. S. M., Bergmann, C., Black, A. K., Brown, A., Carbajal, M. J., ... Wermelinger, S. (2021). A multilab study of bilingual infants: Exploring the preference for infant-directed speech. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1–30. doi: 10.1177/2515245920974622

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, *61*, 1584–1595.

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2022). A systematic review and bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, *7*, 114-133. doi: 10.1038/s41562-022-01452-1

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

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, *8*(2), 181–195. Retrieved 2023-05-31, from `https://linkinghub.elsevier.com/retrieve/pii/S0163638385800059` doi: 10.1016/S0163-6383(85)80005-9

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.

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.

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

Lewis, M., Mathur, M., VanderWeele, T., & Frank, M. C. (2022). The puzzling relationship between multi-lab replications and meta-analyses of the rest of the literature. *Royal Society Open Science*, *9*(2), 211499.

Mathur, M. B., & VanderWeele, T. J. (2020a). Robust metrics and sensitivity analyses for meta-analyses of heterogeneous effects. *Epidemiology*, *31*(3), 356–358.

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

Mathur, M. B., & VanderWeele, T. J. (2021). Meta-regression methods to characterize evidence strength using meaningful-effect percentages conditional on study characteristics. *Research Synthesis Methods*, *12*(6), 731-749.

Schachner, A., & Hannon, E. E. (2011). Infant-Directed Speech Drives Social Preferences in 5-Month-Old Infants. *Developmental Psychology*, *47*(1), 19–25. doi: 10.1037/a0020740

Schreiner, M. S., Zettersten, M., Bergmann, C., Frank, M. C., Fritzsche, T., Gonzalez-Gomez, N., ... Lippold, M. (2022, December). *Limited evidence of test-retest reliability in infant-directed speech preference in a large pre-registered infant sample.* PsyArXiv. Retrieved 2023-06-06, from `https://psyarxiv.com/uwche/` doi: 10.31234/osf.io/uwche

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. doi: 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.

van Rooijen, R., Bekkers, E., & Junge, C. (2019). Beneficial effects of the mother's voice on infants' novel word learning. *Infancy*, *24*(6), 838–856.

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)