Adjudicating discrepancies between large-scale multisite replications and published meta-analyses by accounting for moderators and heterogeneity: A case study on infant-directed speech preference

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

[1]Max Planck Institute for Psycholinguistics
[2]Department of Psychology, Princeton University
[3]Department of Psychology, University of Manitoba
[4]Stanford University
[5]Department of Psychology, University of Oslo
[6]Psychology Department and Neuroscience Center, Brigham Young University
[7]Carnegie Mellon University
[8]Aarhus University
[9]Concordia University
[10]University of British Columbia
[11]Quantitative Sciences Unit, Stanford University

*Author order outside of first and last position is reverse alphabetical.

†Correspondence to: Maya B. Mathur (mmathur@stanford.edu), Quantitative Sciences Unit, 1701 Page Mill Road, Palo Alto, CA, 94304.

Abstract

XXX

Key words: XXX

1. Introduction

Outline:

- MAs are considered the best available source of evidence
- emerging failures to replicate effects thought to be robust based on MAs are first indicators that this might not be the case
- Particularly robust replications are large-scale, multi-site projects (Multi-Lab Replicatins, MLRs)
- MAs are also subject to bias, intransparency, lack of reproducibility - So are they maybe not the best soruce of evidence after all?
- When comparing published MAs and MLRs, Kvarven et al found discrepancies beyond publication bias.
- Lewis et al response
- However, might the summary effect focus be misguided?
- We can learn much more from both meta-analyses and large-scale studies than just whether an effect is real or not
- MAs contain a lot of variation along potential dimensions of interest, such as population, stimuli, and experimental procedure.
- MLRs typically keep the latter two constant and vary mainly the former, this assessing generalizability across specific locations and populations
- MB1 included variation in experimental procedure but not stimuli, although those differed in their relation to participants' past experience outside the lab
- We can thus assess the theoretically important impact of moderators along three dimensions crucial in determining generalizability and boundary conditions for our effect of interest
- We also take on a more sensitive statistical approach

We addressed three primary questions of interest: (1) To what extent do the meta-analysis (MA) and many-labs replication study (MLR) results differ systematically with respect to their average effect sizes? (2) To what extent do the MA and MLR results differ systematically with respect to evidence strength for meaningfully large effects in individual studies? (3) To what extent might study-level moderators account for any such discrepancies? Secondarily, we assessed whether discrepancies in average effect sizes would disappear/decrease when correcting for publication bias in the MA.

1.1. Case study: Infant-directed speech preference and its moderators

A preference for infant- over adult-directed speech, short IDS and ADS,

[introduce the phenomenon]

1.1.1 ManyBabies1

In ManyBabies1: Infant-directed Speech Preference **?**, 69 laboratories on four continents (Asia, Australia, Europe, North America) collected data from over 2000 infants. The main comparison of interest was between auditory stimuli of mothers the infant did not know, who either talked to their own child or to an adult.

[expand with a summary of the approach and main results]

1.1.2 Brief theoretical motivation of moderators

Our moderators included in the confirmatory analyses vary in their theoretical importance, we discuss them in the order they are thought to affect infant performance.

First, age is a key factor. Infants become more mature language processors and accumulate language experience, which allows them to "tune in" to their native language. Note that both the large-scale replication and the meta-analysis report age effects.

Second, whether or not stimuli were presented in infants' native language is likely to affect the strength of their preference (see The ManyBabies Consortium (2020)), although previous work not included in the meta-analysis presented non-native stimuli in an attempt to show that IDS preference is universal and did find that infants prefer IDS even in a non-native language. However, ManyBabies1 reports an effect of nativeness, suggesting that IDS preference may be reduced when tested with non-native, or non-English, stimuli.

Third, 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 to be present in our data as well, among other factors because the current meta-analysis was part of the pooled datasets for the just-cited meta-meta-analysis and because the current replication found method effects. But whether these effects are consistent across datasets is unknown. We group method as follows: Central fixation (cf; including single-screen and eyetracking), Headturn Preference Procedure (hpp), Other (Forced Choice, fc; Conditioned Headturn, cht), because of similarity in the tasks, i.e. either looking to vs away from a single screen with an unrelated visual display (central fixation; cf), turning the head to the side towards flashing lights (headturn preference procedure; hpp), or other tasks which have only been used for a handful of estimates, each (forced choice and conditioned headturn, both being used for 4 estimates, respectively). We will also explore in follow-up analysis whether method effects interact with age, but do not include interactions in the planned analyses due to power concerns.

Fourth, the conditions under which the stimuli were recorded was reported to affect evidence strength in the meta-analysis Dunst et al. (2012). While we overall assume infants prefer infant-directed speech, we expect that the strength of the effect is highest for infant-directed speech produced by a live speaker in the presence of infants, followed by simulated infant-directed speech (i.e. talking 'as if to an infant' during a recording session), with filtered and synthesized speech showing smaller effects in turn. Again, this effect might interact with age, but we do not include this interaction in the preregistered analyses because of power concerns.

Fifth, we expect a stronger effect for a highly familiar speaker, the own main caregiver. In the context of the included studies, this was always the infant's mother. An advantage for maternal speech was reported by van Rooijen et al. (2019) and Barker & Newman (2004). Note that Dunst et al. (2012) distinguish by whether the speaker was a mother talking to her own child, but not whether she was familiar to the participant, versus 'unfamiliar speaker' – but this distinction overlaps with the recording conditions, i.e. whether they were naturalistic or not and thus we opted here for a different, but also theoretically motivated distinction.

Sixth, we also tracked whether infants were presented with an unrelated visual stimulus or saw a video of a speaker, as this methodological variation might heighten their attention in the case of synchronous multi-modal stimuli. Whether this effect leads to an overall longer looking time across conditions (which would not be reflected in the effect size, here a standardized mean difference; SMD) or emphasizes the expected difference between conditions (i.e. a larger SMD) remains to be seen.

Seventh, we investigated the type of dependent variable. Studies either measured infants' looking time to a visual display (related or unrelated to the speech stimuli, i.e. a looming circle or a face), infants' facial expression (e.g. smiling), or their preference for a target. All these measures come with different affordances, and thus might impact the measured effect. Following the meta-analysis, we group the dependent variables into preference (collapsing over looking times and overt responses) and affect.

Finally, we coded whether infants' preference for infant- over adult-directed speech 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 infant-directed speech influences infants' preferences for specific speakers. 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 stimuli. Since most stimuli are not available to us for direct comparison, we use the variable whether IDS preference was a key research question as proxy.

2. Methods

2.1. Preregistration approach

All confirmatory analyses were preregistered prior to data analysis. We wrote the preregistration protocol after we had accessed the meta-analysis dataset (a digitized version of a published table), assembled the replication data (via a public Github repository), and conducted basic cleaning on both datasets, but before conducting any analyses relevant to the research questions herein. (We have written the analysis sections of the preregistration in the past tense to facilitate writing the eventual manuscript even though the analyses had not been conducted at the time of this writing.) 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 meta-analysis data and the replication data) in which the point estimates and their variances had been randomly permuted across the meta-analysis and replications. 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 meta-analysis and replication.

2.2. Meta-analysis dataset

Search strategy From original paper (Dunst et al., 2012): "Studies were located using motherese or parentese or fatherese or infant directed speech or infant-directed speech or infant directed talk or child directed speech or child-directed speech or child directed talk or child-directed talk or baby talk AND infant* or neonate* or toddler* as search terms. Both controlled-vocabulary and natural-language searches were conducted (Lucas & Cutspec, 2007). Psychological Abstracts (PsychInfo), Educational Resource Information Center (ERIC), MEDLINE, Academic Search Premier, CINAHL, Education Resource Complete, and Dissertation Abstracts International were searched. These were supplemented by Google Scholar, Scirus, and Ingenta searches as well as a search of an extensive EndNote Library maintained by our Institute. Hand searches of the reference sections of all retrieved journal articles, book chapters, books, dissertations, and unpublished papers were also examined to locate additional studies. Studies were included if the effects of infant-directed speech on child behavior were compared to the effects of adult-directed speech on child behavior. Studies that intentionally manipulated word boundaries (e.g., Hirsh- Pasek et al., 1987; Nelson, Hirsh-Pasek, Jusczyk, & Cassidy, 1989) or used nonsense words or phrases (e.g., Mattys, Jusczyk, Luce, & Morgan, 1999; Thiessen, Hill, & Saffran, 2005) were excluded." Note that the time span considered is not explicitly mentioned, but the publication year of the meta-analysis is 2012 and the most recent paper in the dataset has been published in 2009.

Adding moderators To supplement the meta-analysis with moderators which were relevant for the research questions in this study but not reported on in the meta-analysis Dunst et al. (2012), it was necessary to re-examine the papers reporting on the original experiments. The added variables were: (1) infants' native language; (2) method; and (3) stimulus language. During this process, a number of possible issues with the original meta-analysis became apparent, such as possibly duplicated effect sizes (from a conference and journal paper of the same authors). Because our goal was to compare the meta-analysis as it was actually conducted and because identifying and correcting apparent errors by the original meta-analysts would introduce substantial subjectivity, we analyzed the dataset exactly as it appeared in the meta-analysis.

We also included raw statistics where possible for future re-computation of effect sizes in a consistent way, but again chose to work with the published effect sizes, which we could not always reproduce.

The meta-analysis comprised $k =$ studies contributing a total of $m =$ estimates, which used a median of $n =$ participants.

2.3. Multisite replication dataset

The replications were conducted by 70 labs, each of which could contribute multiple point estimates for different age groups. The replication dataset comprised $k =$ labs contributing a total of $m =$ estimates, because single labs could contribute data in multiple age groups. (We will use "estimate" to refer to a single point estimate from either the meta-analysis or the replications.) The replications in this dataset had a median of $n =$ participants.

2.3.1 Sampling

Over the course of 14 months, labs were asked to test infants in up to 4 age groups (3-6, 6-9, 9-12, 12-15 months of age) and contribute at least 10 infants to the final sample. To be included, infants had to contribute at least one trial per condition (there were two conditions: infant- and adult-directed speech). Participants were tested across four continents (North America, Europe, Asia, Australia), and grew up learning 12 different languages, and therein 4 different dialects of English; which two being classified as North-American (Canadian and US English) and two as non-North American English (Australian and British English). Since the stimuli were in North American English, they were considered as native only for North American English learning participants. All participants were monolingual; we are not including the data from the bilingual sample sister project (**?**) in the main analyses. Participant exclusion criteria included: (1) out of age range; (2) a known developmental delay; (3) premature birth (before 37 weeks); (3) experimenter error; (4) no usable trials (at least 1 trial per condition with at least 2s total looking time to the screen). Trials were excluded (1) when the minimal looking criterion of 2s was not met; (2) due to technical errors; and (3) because of parental interference. Trials were additionally excluded due to interference, technical error, or infant inattention, as indicated by looking times below 2s. Our dataset has all the 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).

Data extraction The data were downloaded from the public github repository (https://github.com/manyb analysis-public) of the multi-site replication project (The ManyBabies Consortium, 2020). We chose to download difference scores, i.e. they contained per participant the difference in total looking time per trial across the two conditions. This file was the basis for the meta-analysis reported in The ManyBabies Consortium (2020). Effect sizes were computed as standardized mean differences (SMD) based on the average looking time difference divided by its standard deviation on the level of study (i.e. an age group within a lab). Variance was computed accordingly.

2.4. Hypothesized estimate-level moderators

We use "estimate" to refer to a single point estimate corresponding to one experiment within an article in the meta-analysis or alternatively to one experiment consisting of a single age group from one replicating lab (labs in the large-scale replication could test up to four age groups, which count as separate estimates within one study; comparable to an experiment in an article). We use "study" to refer to either one article in the meta-analysis or to one set of replication estimates produced by one replicating lab.

We investigated 8 hypothesized estimate-level moderators of the IDS preference effect, which we coded for all estimates in both the meta-analysis and replication dataset (for an overview see Table 1). These comprised, in addition to source, 1 characteristic of the study population (average participant age [in days, mean-centered]), 4 characteristics of the stimuli (matching the native language of the participant, or not, speech type [natural, filtered, simulated, or synthesized], speaker [own mother or unknown speaker], and mode of presentation [audio or audiovisual]), 2 methodological characteristics (method [Central fixation (cf; including single-screen and eyetracking), Headturn Preference Procedure (hpp), Other (Forced Choice, fc; Conditioned Headturn, cht)], and dependent variable [a type of preference, measured by looking or overt behavior, or affect, e.g. smiling responses of the participants]), and an overall estimate characteristic, namely whether infant-directed speech preference was the main question of a given study or not (binary). One additional factor, infants' native language, was heavily skewed towards English and is confounded with whether stimuli were presented in infants' own native language, as any non-native stimuli were English. We thus leave this factor for exploratory analyses, but mention it here for completeness. For analysis purposes, 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 meta-analysis.

2.4.1 Brief theoretical motivation of moderators

Our moderators included in the confirmatory analyses vary in their theoretical importance, we discuss them in the order they are thought to affect infant performance.

First, age is a key factor. Infants become more mature language processors and accumulate language experience, which allows them to "tune in" to their native language. Note that both the large-scale replication and the meta-analysis report age effects.

Second, whether or not stimuli were presented in infants' native language is likely to affect the strength of their preference (see The ManyBabies Consortium (2020)), although previous work not included in the meta-analysis presented non-native stimuli in an attempt to show that IDS preference is universal and did find that infants prefer IDS even in a non-native language. However, ManyBabies1 reports an effect of nativeness, suggesting that IDS preference may be reduced when tested with non-native, or non-English, stimuli.

Third, 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 to be present in our data as well, among other factors because the current meta-analysis was part of the pooled datasets for the just-cited meta-meta-analysis and because the current replication found method effects. But whether these effects are consistent across datasets is unknown. We group method as follows: Central fixation (cf; including single-screen and eyetracking), Headturn Preference Procedure (hpp), Other (Forced Choice, fc; Conditioned Headturn, cht), because of similarity in the tasks, i.e. either looking to vs away from a single screen with an unrelated visual display (central fixation; cf), turning the head to the side towards flashing lights (headturn preference procedure; hpp), or other tasks which have only been used for a handful of estimates, each (forced choice and conditioned headturn, both being used for 4 estimates, respectively). We will also explore in follow-up analysis whether method effects interact with age, but do not include interactions in the planned analyses due to power concerns.

Fourth, the conditions under which the stimuli were recorded was reported to affect evidence strength in the meta-analysis Dunst et al. (2012). While we overall assume infants prefer infant-directed speech, we expect that the strength of the effect is highest for infant-directed speech produced by a live speaker in the presence of infants, followed by simulated infant-directed speech (i.e. talking 'as if to an infant' during a recording session), with filtered and synthesized speech showing smaller effects in turn. Again, this effect might interact with age, but we do not include this interaction in the preregistered analyses because of power concerns.

Fifth, we expect a stronger effect for a highly familiar speaker, the own main caregiver. In the context of the included studies, this was always the infant's mother. An advantage for maternal speech was reported by van Rooijen et al. (2019) and Barker & Newman (2004). Note that Dunst et al. (2012) distinguish by whether the speaker was a mother talking to her own child, but not whether she was familiar to the participant, versus 'unfamiliar speaker' – but this distinction overlaps with the recording conditions, i.e. whether they were naturalistic or not and thus we opted here for a different, but also theoretically motivated distinction.

Sixth, we also tracked whether infants were presented with an unrelated visual stimulus or saw a video of a speaker, as this methodological variation might heighten their attention in the case of synchronous multi-modal stimuli. Whether this effect leads to an overall longer looking time across conditions (which would not be reflected in the effect size, here a standardized mean difference; SMD) or emphasizes the expected difference between conditions (i.e. a larger SMD) remains to be seen.

Seventh, we investigated the type of dependent variable. Studies either measured infants' looking time to a visual display (related or unrelated to the speech stimuli, i.e. a looming circle or a face), infants' facial expression (e.g. smiling), or their preference for a target. All these measures come with different affordances, and thus might impact the measured effect. Following the meta-analysis, we group the dependent variables into preference (collapsing over looking times and overt responses) and affect.

Finally, we coded whether infants' preference for infant- over adult-directed speech 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 infant-directed speech influences infants' preferences for specific speakers. 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 stimuli. Since most stimuli are not available to us for direct comparison, we use the variable whether IDS preference was a key research question as proxy.

2.5. Measures of evidence strength

We used three statistical metrics to characterize evidence strength for an IDS preference in each source (i.e., the meta-analysis and the multi-lab replication). First, we estimated the mean SMD in each source. Second, we estimated the percentage of population effects[a] in each source that were positive, representing any preference for IDS regardless of magnitude (Mathur & VanderWeele, 2020c, 2019). Third, for a more stringent assessment, we estimated the percentage of population effects in each source representing only effects that might be considered to be meaningfully large (i.e., $SMD > 0.2$) (Mathur & VanderWeele, 2020c, 2019). We use $\widehat{P}_{>0}$ and $\widehat{P}_{>0.2}$ respectively to refer to these estimated percentages of positive and of meaningfully large effects. We compared evidence strength for an IDS preference between sources by estimating differences in each of these three metrics, as detailed below. [Throughout, we will estimate the percentage metrics only if the estimated heterogeneity $\widehat{\tau}$ in a given model is greater than 0.]

2.6. Statistical analyses

2.6.1 Between-source discrepancies before and after accounting for hypothesized moderators

We conducted all statistical analyses in R (R Core Team, 2020).[b] We used robust meta-regression (Hedges et al., 2010; Tipton, 2015) to estimate between-source differences in average effect sizes, in the percentages of positive effects (Mathur & VanderWeele, 2020b). These methods are similar to generalized estimating equations in that they accommodate potential correlation between point estimates contributed by the same paper or lab and obviate the distributional assumptions that would be required by parametric multilevel modeling. Specifically, we fit two meta-regression models:[c] (1) a **naïve model** that compared the two sources but did not account for other hypothesized moderators and (2) a **moderated model** that additionally included the other hypothesized moderators. These two models estimated the extent to which the meta-analysis and replication results differed when either ignoring estimate-level moderators (the naïve model) or when accounting for them (the moderated model). Both models included all $m = 155$ estimates from both data sources.

First, we fit a **naïve model** in which the only meta-regressive covariate besides the intercept was source (meta-analysis versus multi-lab replication). The coefficient of source ($\widehat{\beta}_{\text{naïve}}$) estimated how much larger, on average, effect sizes were in the meta-analysis compared to the replication project, while ignoring the hypothesized estimate-level moderators (specifically, when averaging over their distributions across sources). We also used this model to robustly estimate the percentage of positive effects and of meaningfully strong effects in each source (Mathur & VanderWeele, 2020b). To estimate inference for these percentages while accounting for possible correlation of estimates within studies, we resampled studies with replacement (Davison & Hinkley (1997), Chapter 3.8) and constructed confidence intervals using the bias-corrected and accelerated method (Carpenter & Bithell, 2000; Efron, 1987; Mathur & VanderWeele, 2020b). We also estimated the difference between these percentages for the meta-analysis versus the replication project, which we call $\Delta_{\text{naïve}}\left(\widehat{P}_{>0}\right)$ and $\Delta_{\text{naïve}}\left(\widehat{P}_{>0.2}\right)$,

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

[b] We used the following R packages: XXX.

[c] Say we also fit subset models to estimate Phats because heterogeneity seemed very different in reps vs. MA

along with inference (Mathur & VanderWeele, 2020b).

Second, we fit a **moderated model** that included covariates corresponding to each of the hypothesized moderators (Section 2.4) as well as the covariate for source, whose coefficient ($\widehat{\beta}_{\text{mod}}$) estimated how much larger, on average, effect sizes were in the meta-analysis compared to the replication project while holding constant all hypothesized moderators. [We anticipate that the model may not be statistically estimable if some moderators are relatively highly correlated, in which case we will remove moderators one-by-one based on ascending order of scientific relevance, as delineated above, until the model is estimable.] We also estimated the average effect size in each source, holding constant all hypothesized moderators to their reference levels (i.e., the most common category in the meta-analysis). We estimated the percentages of positive effects and of meaningfully large effects ($\Delta_{\text{mod}}\left(\widehat{P}_{>0}\right)$ and $\Delta_{\text{mod}}\left(\widehat{P}_{>0.2}\right)$), again while holding constant all hypothesized moderators to their reference levels.

Then, we assessed the extent to which accounting for estimate-level moderators reduced between-source discrepancies by comparing the naïve and moderated model estimates with respect to each metric of evidence strength. That is, we calculated $\widehat{\beta}_{\text{mod}} - \widehat{\beta}_{\text{naïve}}$, the absolute reduction in the between-source difference in average effect sizes upon accounting for the hypothesized moderators, as well as $\Delta_{\text{naïve}}\left(\widehat{P}_{>0}\right) - \Delta_{\text{mod}}\left(\widehat{P}_{>0}\right)$ and $\Delta_{\text{naïve}}\left(\widehat{P}_{>0.2}\right) - \Delta_{\text{mod}}\left(\widehat{P}_{>0.2}\right)$, the absolute reductions in the between-source differences in the percentages of positive and of meaningfully strong effects.

To investigate how much of the heterogeneity in our model was associated with the hypothesized moderators, we followed a two-step procedure. We estimated the residual heterogeneity in each model, $\widehat{\tau}_{\text{naïve}}$ and $\widehat{\tau}_{\text{mod}}$ (Hedges et al., 2010), representing the estimated standard deviation of the population effects that remains after holding constant all covariates in the model. We then calculated $\widehat{\tau}_{\text{naïve}} - \widehat{\tau}_{\text{mod}}$, representing the reduction in residual heterogeneity after accounting for the hypothesized moderators as well as source. We estimated inference for these cross-model comparisons using bias-corrected and accelerated bootstrapping (Efron (1987); Carpenter & Bithell (2000); c.f. Lockwood & MacKinnon (1998) for a similar approach).

2.6.2 Publication bias

We assessed the possible contribution of publication bias to the meta-analysis results and to between-source discrepancies in average effect sizes. First, we assessed publication bias in the meta-analysis using selection model methods (Vevea & Hedges, 1995), sensitivity analysis methods (Mathur & VanderWeele, 2020d), and the significance funnel plot (Mathur & VanderWeele, 2020d). These methods assume that the publication process favors “statistically significant” (i.e., $p < 0.05$) and positive results over “nonsignificant” or negative results, a typically reasonable assumption that also conforms well to empirical evidence on how publication bias operates in practice (Jin et al., 2015; Mathur & VanderWeele, 2020a). “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. We used the sensitivity analysis methods to estimate 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). These methods, unlike the selection model, also accommodated the point estimates' non-independence within articles and did not make any distributional assumptions (Mathur & VanderWeele, 2020d). [If the worst-case estimate in the meta-analysis is less than the estimated mean in the replications, we will also report the amount of publication bias required to shift the estimate in the meta-analysis to match the estimate in the replications (Mathur & VanderWeele, 2020d).]

2.6.3 Sensitivity analyses

A small number of studies in the meta-analyses used between-subjects rather than within-subjects designs. We anticipated that between-subjects designs may differ systematically because larger unintended variation between conditions in between-participant 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 2.6.1 after excluding between-subjects studies.

We visually inspected the linearity assumption for age in the meta-regression model. [If the linearity assumption appears not to hold, we will conduct sensitivity analyses with age broken into categories.]

3. Results

3.1. Between-source discrepancies before and after accounting for hypothesized moderators

Among only the studies from the meta-analysis, the estimated average effect size was 0.70 (95% CI: [0.37, 1.02]; p = 0.0006), with considerable heterogeneity (estimated standard deviation of population effects $\widehat{\tau} = 0.45$). We estimated that nearly all[d] of the population effects were positive (90% [95% CI: 74%, 98%]) and were stronger than $SMD = 0.2$ (86% [95% CI: 70%, 98%]). Among only the replication studies, the estimated average effect size was half as large (0.35 (95% CI: [0.28, 0.43]; p = <0.0001)) and with less heterogeneity ($\widehat{\tau} = 0.11$). Despite the much smaller mean estimate in the replications compared to the meta-analysis, we again estimated that nearly all of the population effects were positive (100% [95% CI: 86%, 100%]) and were stronger than $SMD = 0.2$ (90% [95% CI: 66%, 100%]), similar to the meta-analysis. This occurred because effects in the replication studies appeared to be much more concentrated around the mean than effects in the meta-analysis (**Figure 1**).

[d] Cite MRM when describing this analysis in methods because uses the cluster bootstrap.

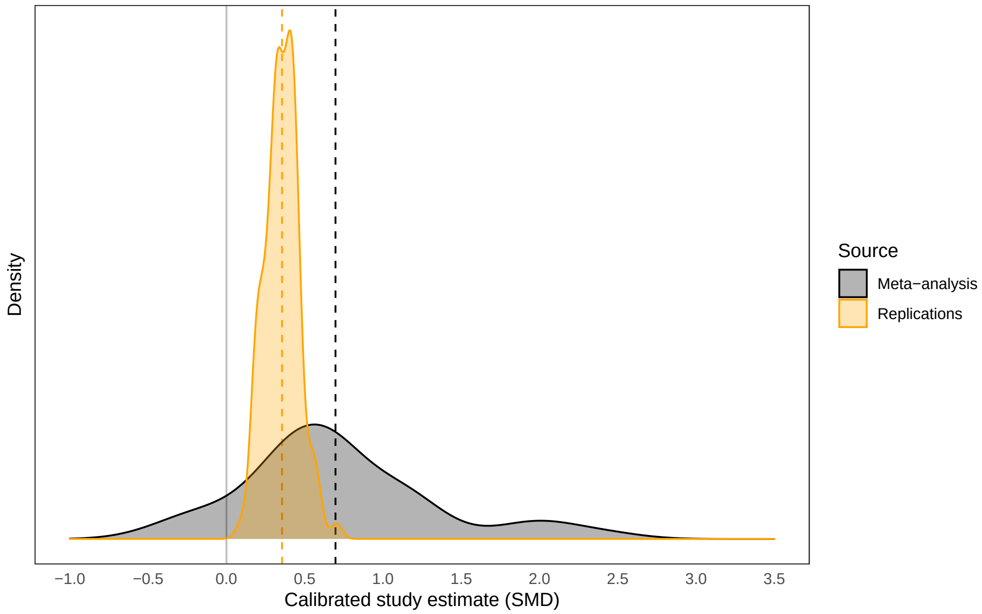

Figure 1: *Estimated densities of population effects in the meta-analysis (orange) and in the replications (gray). Vertical dashed lines: mean estimates from each source. Vertical gray line: null.*

In the **naïve model** combining the two sources, the estimated average effect sizes in the meta-analysis and in the replications, respectively, were $SMD = 0.68$ (95% CI: [0.35, 1]; $p = 8 \cdot 10^{-4}$) and $SMD = 0.35$ (95% CI: [0.28, 0.43]; $p = <0.0001$). Thus, effect sizes in the meta-analysis were larger by on average 0.32 (95% CI: [0, 0.64]; $p = 0.05$) units on the standardized mean difference (SMD) scale. There was considerable residual heterogeneity (estimated standard deviation of population effects $\widehat{\tau}_{\text{naïve}} = 0.35$).

The **moderated model** was estimable with the inclusion of three moderators besides source: mean age, language, and method. **Table 2** shows the estimated associations of each moderator with effect sizes. [Discuss the estimates for each moderator and which seemed most important.] In the moderated model, the estimated average effect size in the meta-analysis and in the replications when setting the moderators to their average values in the meta-analysis, respectively, $SMD = 0.61$ (95% CI: [0.26, 0.96]; $p = 2 \cdot 10^{-3}$) and $SMD = 0.13$ (95% CI: [−0.08, 0.35]; $p = 0.22$). Thus, effect sizes in the meta-analysis were larger by on average 0.48 (95% CI: [−0.02, 0.97]; $p = 0.06$) units on the standardized mean difference scale. This discrepancy that was, if anything, larger than that seen in the naïve model, and the residual heterogeneity appeared essentially unchanged ($\widehat{\tau}_{\text{mod}} = 0.33$). In the meta-analysis and replications respectively, we estimated that XXX% of population effects were positive; and we estimated that XXX% of population effects were stronger than $SMD = 0.2$.

Moderator	EstCI	Pval
X.Intercept.	0.13 [-0.08, 0.35]	0.2174876
isMetaTRUE	0.48 [-0.02, 0.97]	0.0593261
mean_agec	0.02 [0.01, 0.03]	0.0005402
test_langb.nonnative	-0.09 [-0.20, 0.02]	0.1026598
test_langc.artificial	-0.5 [-2.49, 1.48]	0.3881219
methodb.hpp	0.11 [-0.23, 0.46]	0.5047620
methodc.other	0.67 [-1.17, 2.52]	0.2844915

Table 2: *Meta-regressive estimates of moderation by various study design and participant characteristics. Intercept: estimated mean SMD when all listed moderators are set to 0. For binary covariates, estimates represent SMDs for the increase in a study's effect size associated with a study's having, versus not having, the covariate. For mean age, the estimate represents the SMD for the increase in effect size associated with a 1-month increase in mean participant age. Bracketed values are 95% confidence intervals. CI: confidence interval. p-values 889 are for the moderators' coefficients themselves, not for the subset of studies having the listed characteristic.*

Comparing the two models indicated that controlling for the hypothesized moderators in fact *increased* the estimated between-source discrepancy in average effect sizes by 0.15 (95% CI: [0.13, 0.47]) units on the SMD scale and increased the discrepancies in the percentages[e] of positive and of meaningfully large population effects by (95% CI: [,]) and (95% CI: [,]) percentage points respectively.

3.2. Publication bias

The meta-analysis contained 22 affirmative and 29 nonaffirmative studies. The meta-analytic average corrected for publication bias was $SMD = 0.92$ (95% CI: [0.60, 1.23]; p <0.0001; Vevea & Hedges (1995)), which was in fact somewhat larger than the uncorrected estimate of $SMD = 0.68$. The sensitivity analyses for publication bias indicated that 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 meta-analytic average would decrease to 0.21 (95% CI: [−0.06, 0.48]). This worst-case estimate arises from meta-analyzing only the observed "nonsignificant" or negative studies and excluding the observed "significant" and positive studies (Mathur & VanderWeele, 2020d). In order for publication bias to fully explain the discrepancy between the meta-analysis and the replications (i.e., to reduce the meta-analysis estimate to match the estimate of 0 in the replications), affirmative studies would need to be 7.8 times more likely to be published than nonaffirmative studies.[f] As a graphical heuristic, the significance funnel plot in **Figure 2**

[e] still need to code these

[f] For Discussion, could compare this to our estimates from Metalab and psychology more broadlyMathur & VanderWeele (2020d); this amount of publication bias is quite high but maybe within the realm of possibility. Overall, the presence of so many nonaffirmative studies in the meta-analysis, the fact that the selection model did not detect much if any publication bias in the expected direction, and the sensitivity analyses seem to suggest that publication bias is not a sufficient explanation.

relates studies' point estimates to their standard errors and compares the pooled estimate within all studies (black diamond) to the worst-case estimate (grey diamond). When the diamonds are close to one another or the grey diamond represents a positive, nonnegligible effect size, the meta-analysis may be considered fairly robust to publication bias (Mathur & VanderWeele, 2020d).

[Taken together, the results from the selection model and from the sensitivity analysis suggest [...]

Reproducibility

All code, materials, and data required to reproduce this research are publicly available and documented (XXX).

Acknowledgments

This research was funded by a grant from XXX. MBM was supported by a grant from the Franklin E. Fetzer Memorial Trust (award R2020-16). The funders had no role in the design, conduct, or reporting.

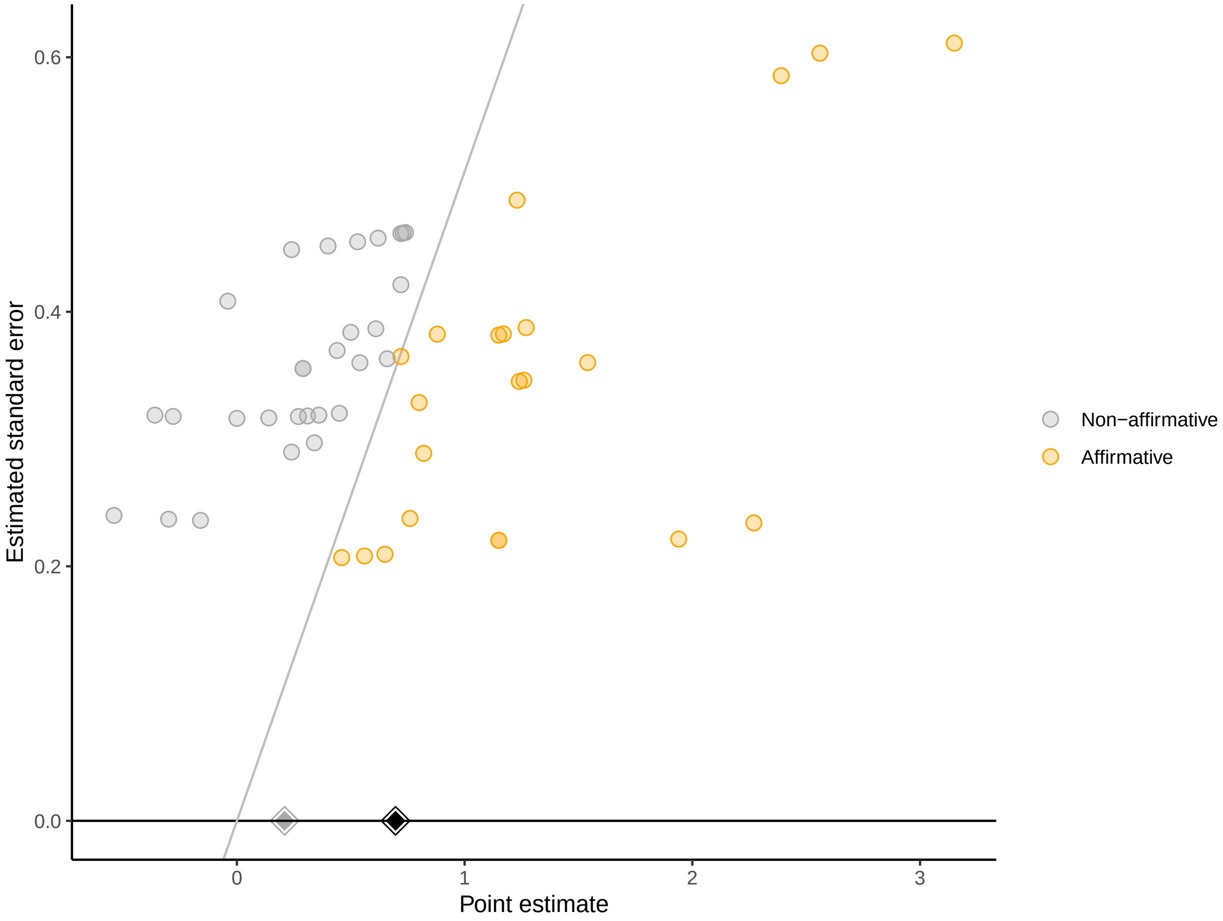

Figure 2: *Significance funnel plot displaying studies' point estimates versus their estimated standard errors. Orange points: affirmative studies ($p < 0.05$ and a positive point estimate). Grey points: nonaffirmative studies ($p \geq 0.05$ or a negative point estimate). Diagonal grey line: the standard threshold of "statistical significance" for positive point estimates; studies lying on the line have exactly $p = 0.05$. Black diamond: main-analysis point estimate within all studies; grey diamond: worst-case point estimate within only the nonaffirmative studies.*

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.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, *19*(9), 1141–1164.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* (No. 1). Cambridge University Press.

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.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171–185.

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.

Jin, Z.-C., Zhou, X.-H., & He, J. (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine*, *34*(2), 343–360. (https://doi.org/10.1002/sim.6342)

Lockwood, C. M., & MacKinnon, D. P. (1998). Bootstrapping the standard error of the mediated effect. In *Proceedings of the 23rd annual meeting of SAS Users Group International* (pp. 997–1002).

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). Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Research Synthesis Methods*. (In press.)

Mathur, M. B., & VanderWeele, T. J. (2020b). Meta-regression estimates of the percentage of meaningfully strong population effects.
(Preprint retrieved from https://osf.io/bmtdq/)

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

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

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

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. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375.

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)

Table 1: *The distribution of moderators in the meta-analysis (MA) and large-scale replication ManyBabies1 (MB).*

	MA	MB	p	test
n	51	104		
study_type = MB (%)	0 (0.0)	104 (100.0)	<0.001	
mean_agec (mean (SD))	0.00 (6.61)	11.78 (7.63)	<0.001	
test_lang = nonnative (%)	0 (0.0)	58 (55.8)	<0.001	
native_lang (%)			0.001	
cantonese	4 (7.8)	0 (0.0)		
dutch	0 (0.0)	5 (4.8)		
english	47 (92.2)	62 (59.6)		
french	0 (0.0)	6 (5.8)		
german	0 (0.0)	16 (15.4)		
hungarian	0 (0.0)	2 (1.9)		
italian	0 (0.0)	1 (1.0)		
japanese	0 (0.0)	4 (3.8)		
korean	0 (0.0)	3 (2.9)		
norwegian	0 (0.0)	1 (1.0)		
spanish	0 (0.0)	2 (1.9)		
swissgerman	0 (0.0)	1 (1.0)		
turkish	0 (0.0)	1 (1.0)		
method (%)			<0.001	
a.cf	34 (66.7)	69 (66.3)		
b.hpp	10 (19.6)	35 (33.7)		
c.other	7 (13.7)	0 (0.0)		
speech_type (%)			<0.001	
a.simulated	28 (54.9)	0 (0.0)		
b.naturalistic	16 (31.4)	104 (100.0)		
c.filtered	4 (7.8)	0 (0.0)		
d.synthesized	3 (5.9)	0 (0.0)		
own_mother = b.yes (%)	4 (7.8)	0 (0.0)	0.019	
presentation = b.video recording (%)	15 (29.4)	0 (0.0)	<0.001	
dependent_measure = b.affect (%)	7 (13.7)	0 (0.0)	0.001	
main_question_ids_preference = b.no (%)	11 (21.6)	0 (0.0)	<0.001	