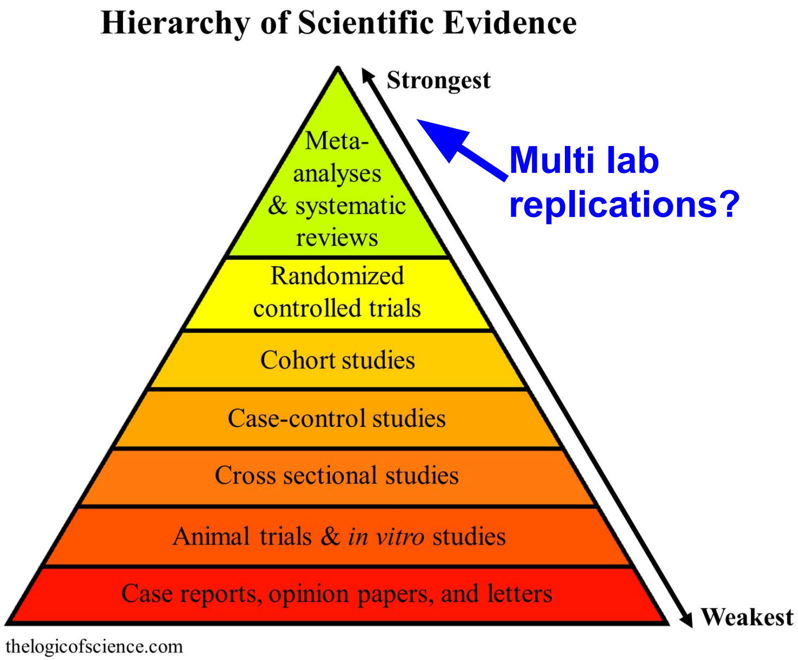
Why do large-scale replications and meta-analyses diverge? A case study of infant-directed speech preference

Molly Lewis
Carnegie Mellon University

Christina Bergmann, Martin Zettersten, Melanie Soderstrom, Angeline Sin Mei Tsui, Julien Mayor, Rebecca A. Lundwall, Jessica E. Kosie, Natalia Kartushina, Riccardo Fusaroli, Michael C. Frank, Krista Byers-Heinlein, Alexis K. Black, and Maya B. Mathur

What’s the best way to estimate the size of important effects in psychology?



Meta-analysis = Statistical aggregation of effects from existing literature

Multi-lab replications = Coordinated replications across many labs

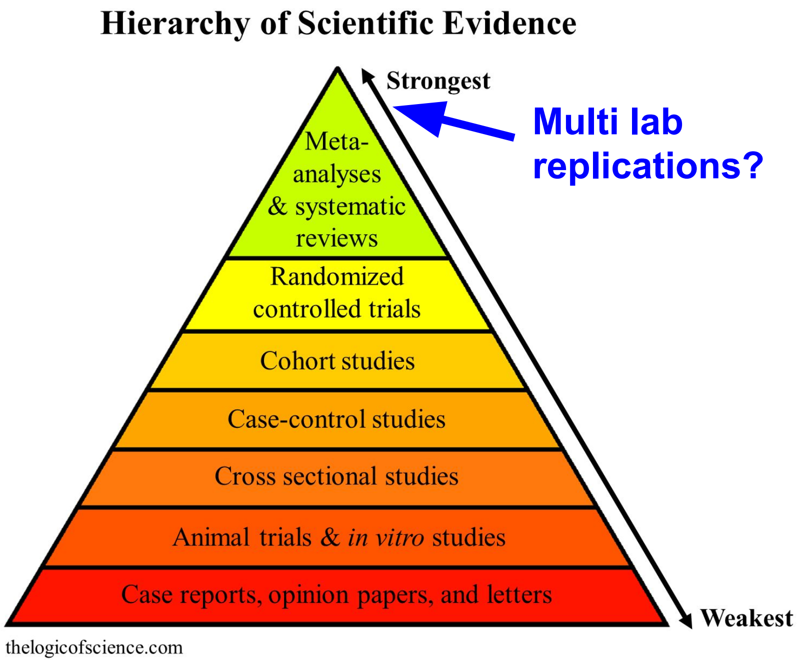
These methods have different strengths/weaknesses

Meta-Analyses:

- Relatively few resources
- Variability in population, stimuli, method
- Individual studies typically not pre-registered; subject to publication bias

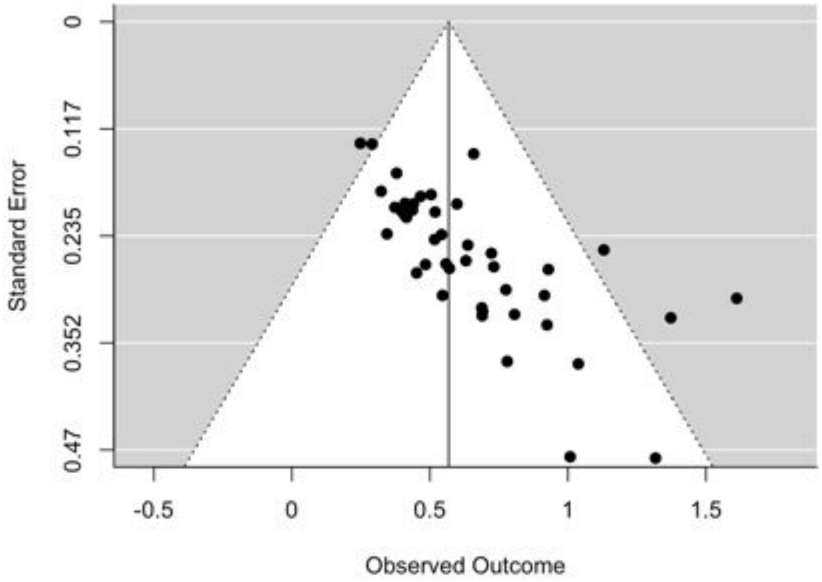
Multi-Lab Replications:

- Highly resource intensive
- Standardization of stimuli and method; some variability in populations
- Typically pre-registered



What’s the relationship between aggregate estimates derived using these two methods?

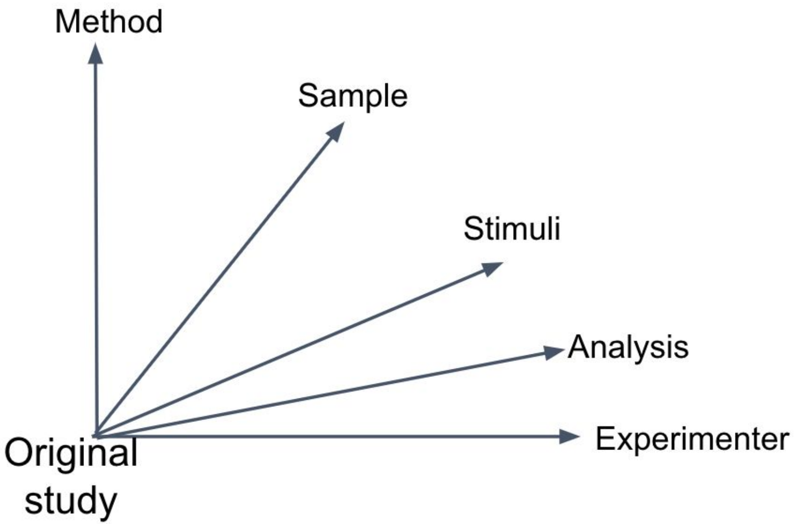
- Naively, expect them to be the same
- But, recent work suggests they are discrepant (Kvarven, et al, 2020)
- ES from MAs three times larger than MLRs
- Due to publication bias?
- Evidence that publication bias can’t fully account for discrepancy (Lewis, et al., 2020)



(Shanks, et al. 2015)

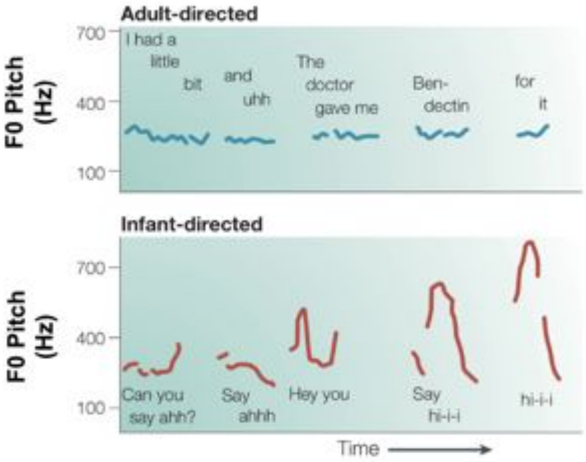
Why the discrepancy? (Lewis et al., 2020)

- Another possibility: Heterogeneity
- MAs contain more heterogeneity along relevant dimensions
- MAs are adapted to their local context, whereas MLRs are typically not
- Perhaps accounting for these moderators will reveal the source of the discrepancy.



Case Study: Infant directed speech preference

Do babies prefer to listen to infant directed speech (IDS), compared to adult directed speech (ADS)?



Kuhl (2004) - originally Fernald & Kuhl (1987)

Shorter utterances, higher, varied pitch, longer pauses

Case Study: Infant directed speech preference

(Source: Moll & Tomasello, 2010)

Dependent measure:
Looking time to checkerboard

Independent variable:
ADS vs. IDS played in pairs of trials within subjects

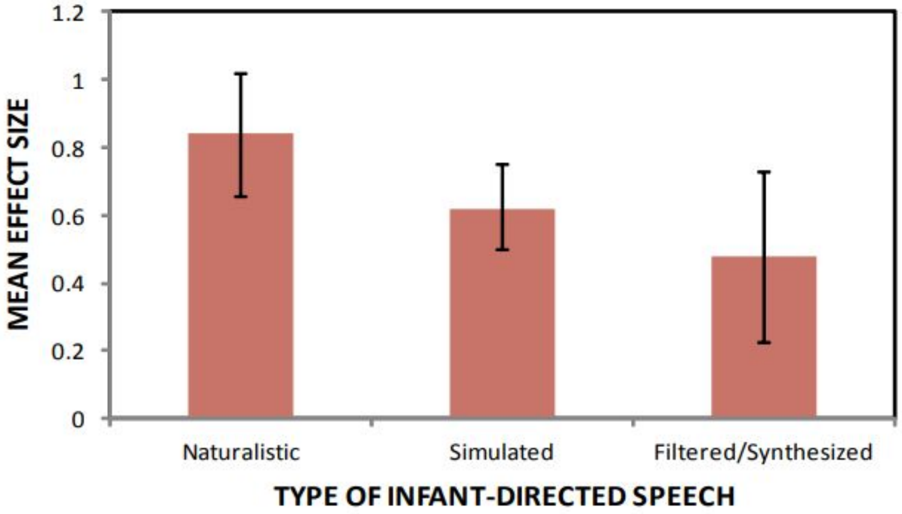
FIG. 2.—Mean looking times (in sec) of 1-month-old subjects from Experiment 1 (including standard errors); ID = infant-directed and AD = adult-directed.

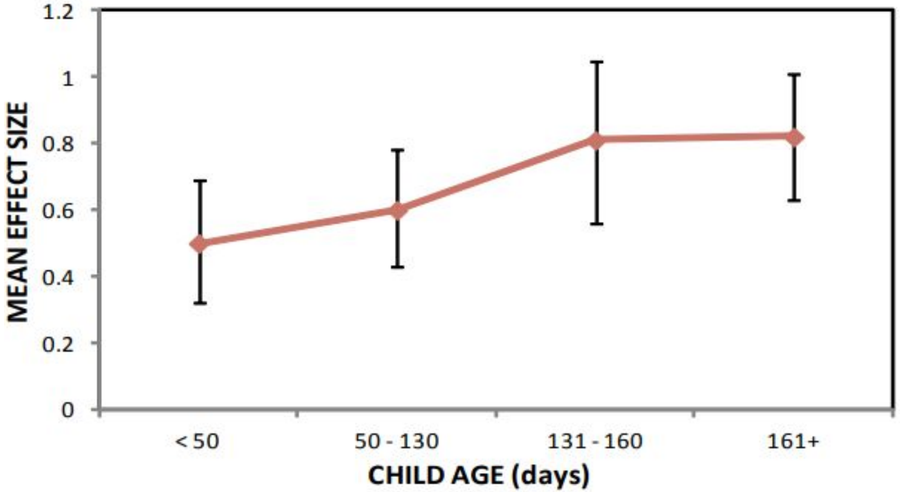
(Cooper & Aslin, 1990)

Effect size

Meta-analysis of IDS preference (Dunst, Gorman, & Hamby, 2012)

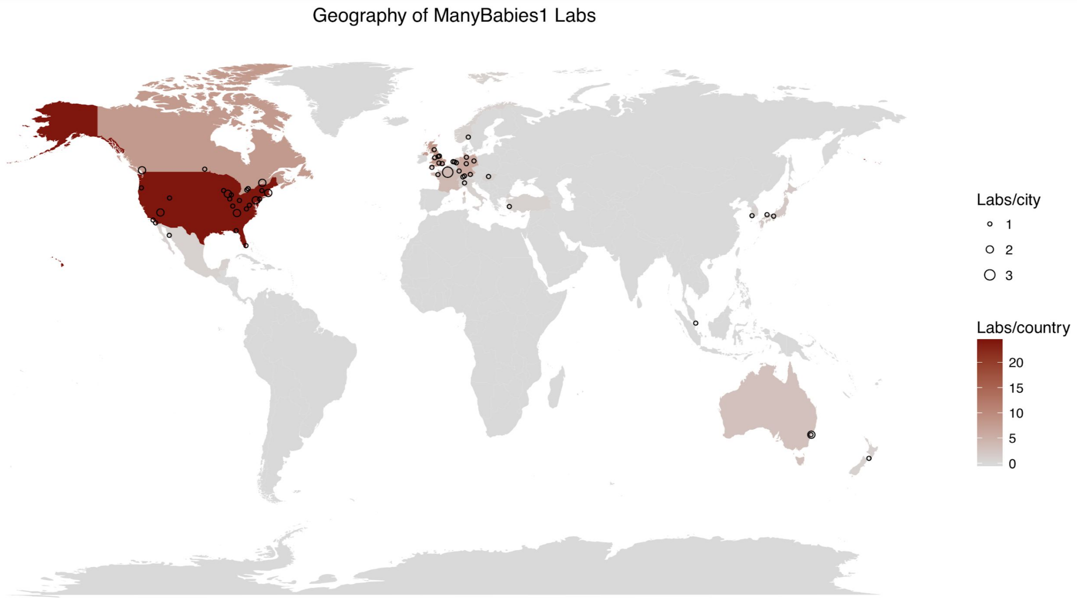
- *N* = 34 studies (840 infants), published 1983-2011
- Aggregate ES = 0.67 (CI = [0.57-0.76])



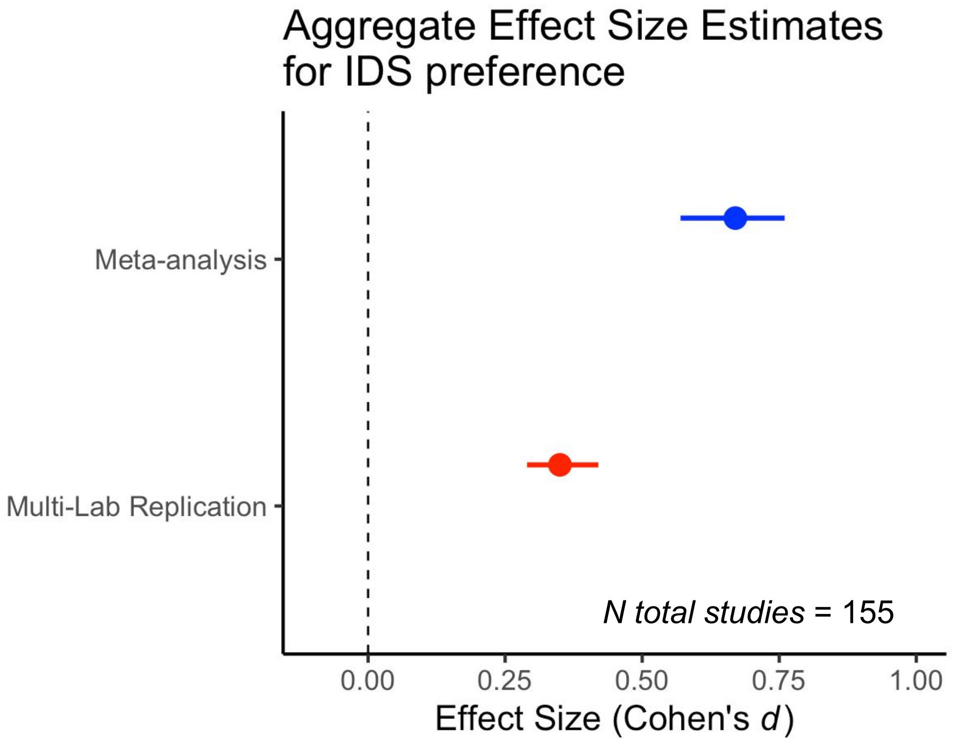


Multi-Lab Replication of IDS preference (ManyBabies, 2020)

- Each lab conducted their own replication based on Cooper & Aslin (1990)
- Consensus design
- 67 labs, 2,329 babies!
- Constant stimuli, DV
- Some variation in method
- Aggregate ES = 0.35 (CI = [0.29-0.41])



The current work



- As found previously, meta-analytic ES > multi-lab ES (discrepancy = 0.32)
- Why?
- Systematically compared effect sizes from two sources, accounting for possible differences due to heterogeneity by coding same set of moderators in each

Moderators we examined for both data sources

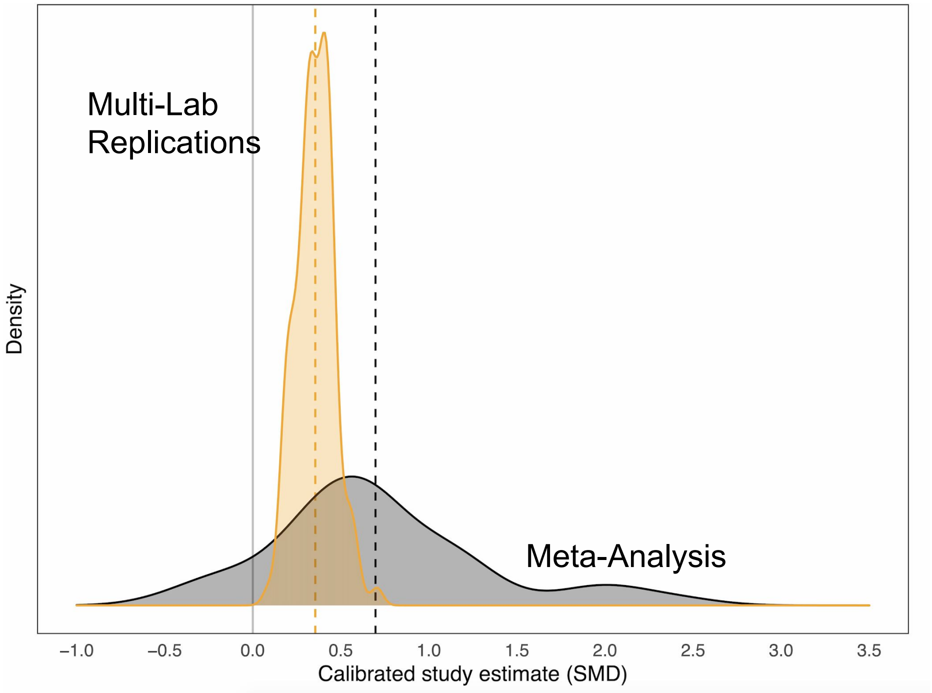
1. Age
2. Test language (native vs. non-native)
3. Method (central fixation vs. headturn preference procedure vs. other)
4. Speech type (Infant directed speech vs. simulated infant directed speech vs. synthesized speech)
5. Speech source (caregiver vs. other)
6. Visual stimulus (unrelated vs. speaker)
7. DV type (looking time vs. facial expression vs. preference for target)
8. Target research question (primary vs. secondary)

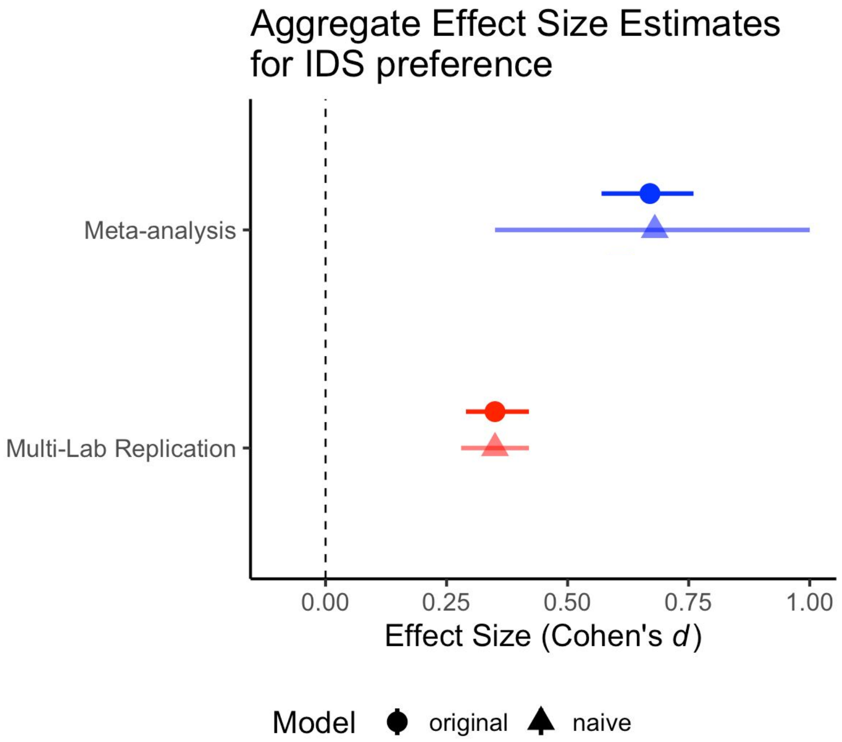
Analysis Approach

- Fit *both* meta-analytic and multi-lab replication data in single meta-analytic model (robust meta-regression; Hedges et al., 2010; Tipton, 2015)
- **Naive model:** Source (MA vs. MLR) as only moderator
- **Moderated model**: Source + 8 moderators that should affect outcomes based on past research (additive)
  - Continuous moderators centered; reference levels for factors defined by most frequent MA level
  - *Model only able to converge with 3 moderators (age, test language, method)
- Planned analyses pre-registered

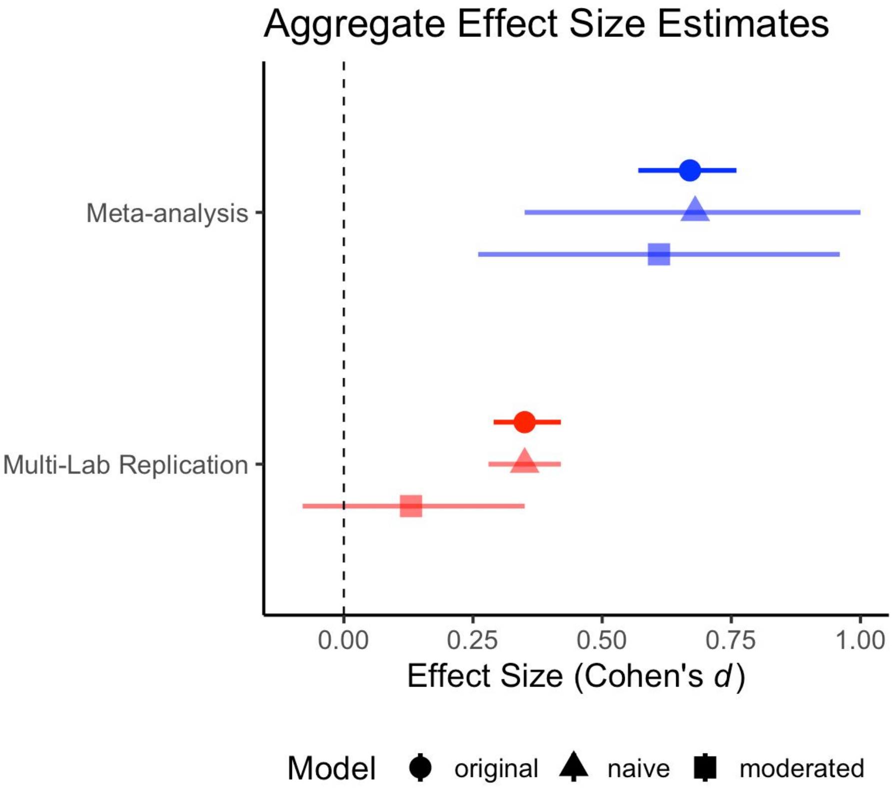
Results: Naive Model

MA - MLR Discrepancy = .32 [0, .64]
Tau = .35





Results: Moderated Model



Moderator	Est [95 CI]	p
intercept	0.13 [-0.08, 0.35]	0.22
is-MA (true)	0.48 [-0.02, 0.97]	0.06
mean age	0.02 [0.01, 0.03]	$<.001$
test language (non-native)	-0.09 [-0.20, 0.02]	0.10
test language (artificial)	-0.5 [-2.49, 1.48]	0.39
method (hpp)	0.11 [-0.23, 0.46]	0.51
method (other)	0.67 [-1.17, 2.52]	0.28

MA - MLR Discrepancy = .48 [-.02, .97]
Tau = .33

Could the discrepancy be due to publication bias in the MA?

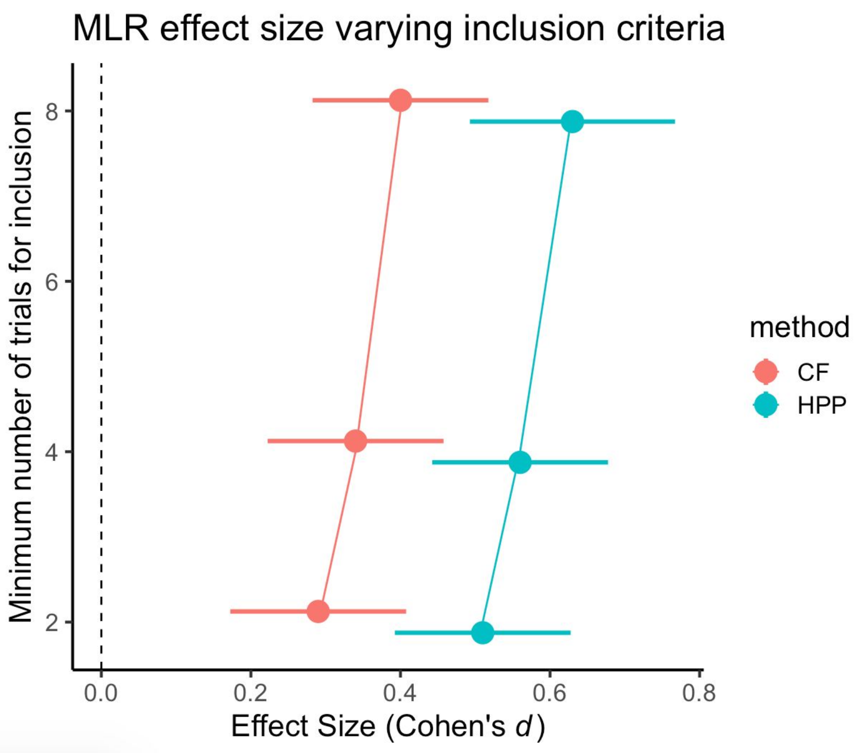
- Probably not…
- After correcting for publication bias (Vevea & Hedges, 1995), the ES was actually larger (.92 CI = [.6-1.23])
- Sensitivity analysis for publication bias (Mathur & VanderWeele, 2020 - see Maya's talk today!)
  - Worst case scenario = "statistically significant" positive results are infinitely more likely to be published than "nonsignificant" or negative results
  - Meta-analyze only non-significant/negative studies
  - Significant studies would have to be about 8 times more likely to be published than nonsignificant/negative studies to eliminate discrepancy

Discussion

- Even when analyzed within the same model and controlling for moderators, MA effect size more than twice as big as MLR effect size
- Probably not due (entirely) to publication bias in MA
- Next: Update MA with recent papers since 2011
- Extend ManyBabies1 dataset with existing or pending spin-off studies
    - ManyBabies1-Bilingual (Byers-Heinlein et al., 2020/in press; 333 participants, 17 labs)
    - Test-retest reliability (Schreiner et al., in prep; 149 participants, 7 labs)
    - ManyBabies1-Africa (Tsui et al., in prep; data collection planned for 2021-2022)
    - Native language follow-up (7 labs signed up; data collection ongoing)

Other possible sources of discrepancy

- Still lots of residual heterogeneity - look at other moderators (e.g., by fitting separate models)
- Difference in inclusion criteria between ManyBabies and MA
- Others?



Thanks!

Papers:

Pre-registration: https://osf.io/scg9z

Lewis, Mathur, VanderWeele, & Frank (2020): https://psyarxiv.com/pbrdk

Mathur & VanderWeele (2020, *J. Royal Stat. Society: Series C*): https://osf.io/s9dp6/

IDS MLR (ManyBabies; 2020, *AMPPS*): https://psyarxiv.com/s98ab

mollyllewis@gmail.com | mllewis | mollyllewis

Appendix

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	

Table 1

Average Weighted Cohen's d and 95% Confidence Intervals for Different Speech Conditions

Condition	Number: Studies	Number: Effect Sizes	Average Effect Size	95% Confidence Intervals	*Z*	*p*-value
Speaker						
Mothers	20	30	0.61	0.48-0.74	8.97	.0000
Unfamiliar Adults	14	21	0.73	0.58-0.87	10.06	.0000
Speech Presentation						
Audio Recordings Only	26	36	0.62	0.51-0.73	11.14	.0000
Audio + Video	8	15	0.82	0.61-1.03	7.67	.0000
Child Outcome						
Preference Measure	33	44	0.64	0.54-0.75	12.33	.0000
Positive Affect	7	7	0.87	0.56-1.18	5.49	.0000

Table 2

Moderator Analyses of the Relationship Between Infant-Directed Speech and the Child Preference Measures

Moderators	Number: Studies	Number: Effect Sizes	Average Effect Size	95% Confidence Intervals	*Z*	*p*-value
Year of Publication						
< 1991	13	16	0.92	0.72-1.09	10.38	.0000
1991 – 1995	12	20	0.56	0.41-0.72	7.09	.0000
1995 +	9	15	0.53	0.35-0.71	5.83	.0000
Type of Design						
Between Conditions	29	42	0.71	0.60-0.81	12.87	.0000
Between Group	5	9	0.49	0.26-0.71	4.19	.0000
Type of Study						
Journal Article	33	49	0.66	0.55-0.76	12.87	.0000
Other	1	2	0.84	0.42-1.26	3.92	.0001
Setting						
Child's Home	2	2	2.47	1.65-3.29	5.88	.0000
Laboratory	32	49	0.64	0.54-0.72	12.82	.0000