

Intro to Meta-Analysis

23 March 2020

Modern Research Methods

Repeating an experiment

- “Replication” – core tenet of science
- Many examples (Zettersten & Lupyan, 2020, IDS vs. ADS, Mutual exclusivity, Vohs & Schooler, 2017)
- In each of these cases, we want to know how one or more replications compare to each other.
- Discussed several statistical tools for evaluating this (p-values, confidence intervals, effect sizes)
- In many cases in psychology, we see failed replications (and we talked about tools for fixing that, e.g. pre-registration)
- So far, we’ve just compared outcomes for two experiments – but what if we have many replications?

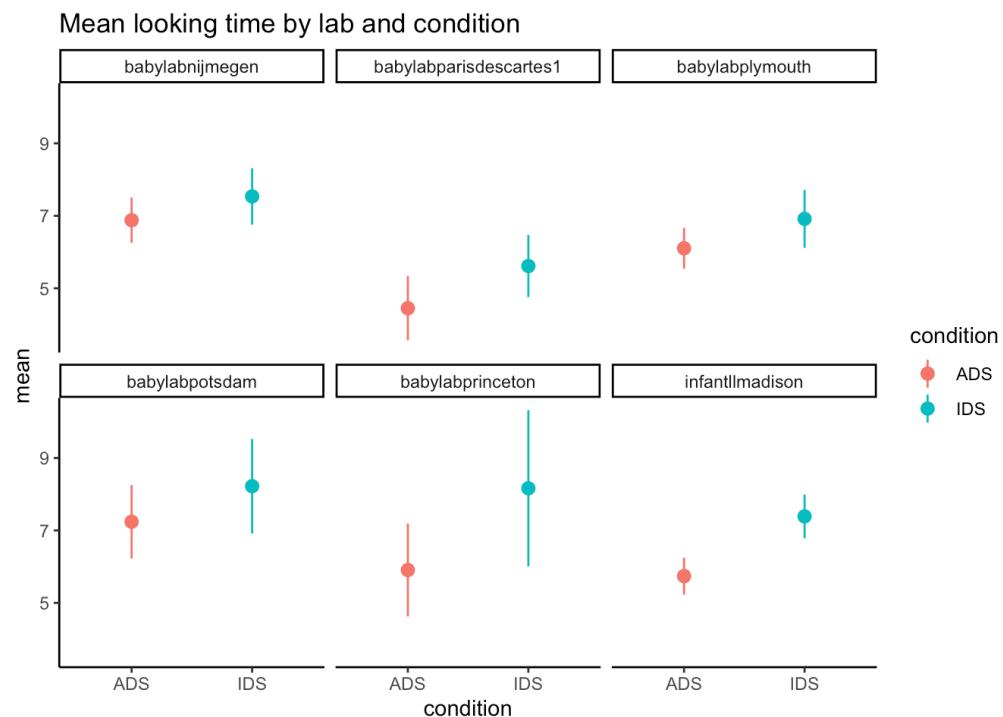
Overview of course

- 1) Philosophy of Cumulative Science
- 2) The Single Experiment – Experimental design, tools in R for working with data and plotting data, reproducibility
- 3) Repeating an Experiment – Intro to statistical concepts, replication of experiments
- 4) Aggregating Many Experiments – Meta-analysis

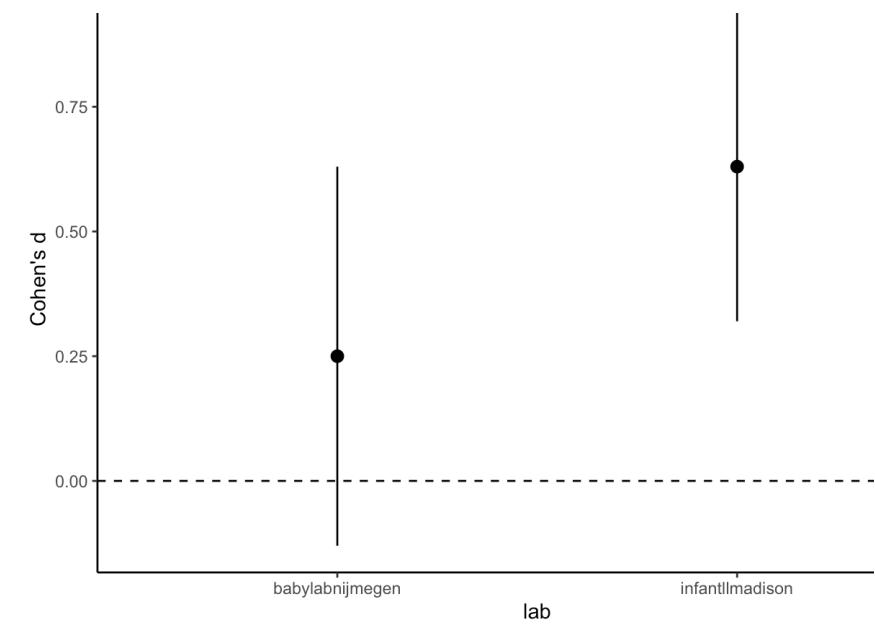
A repeated experiment



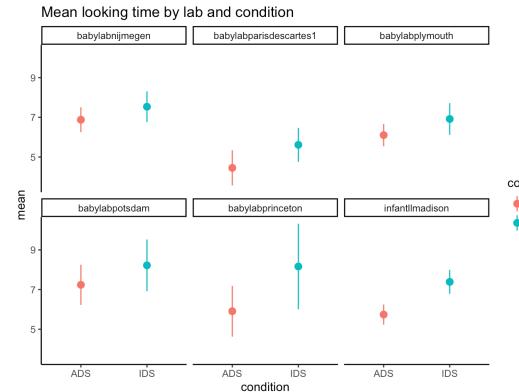
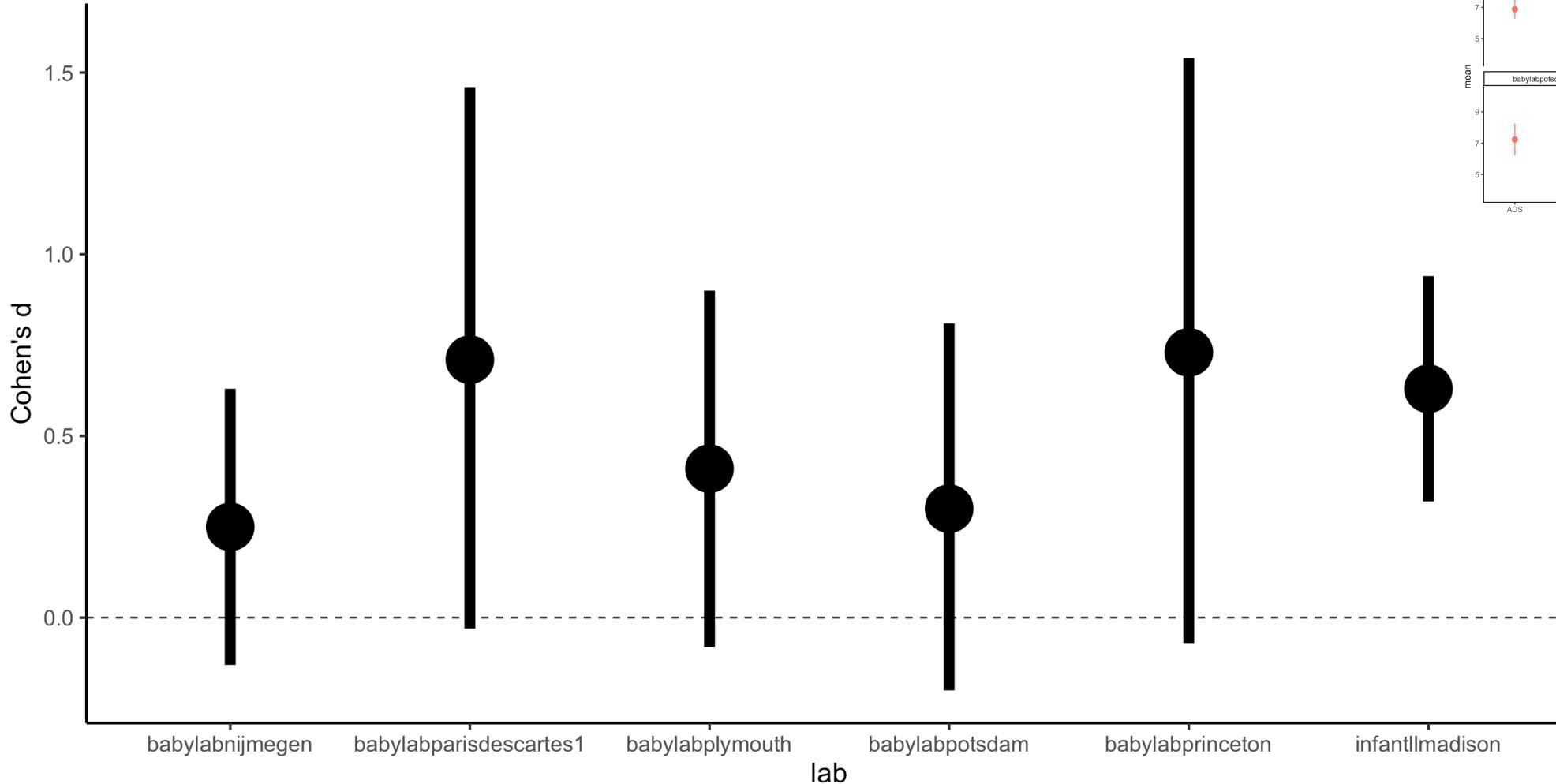
Assignment 5, 9b



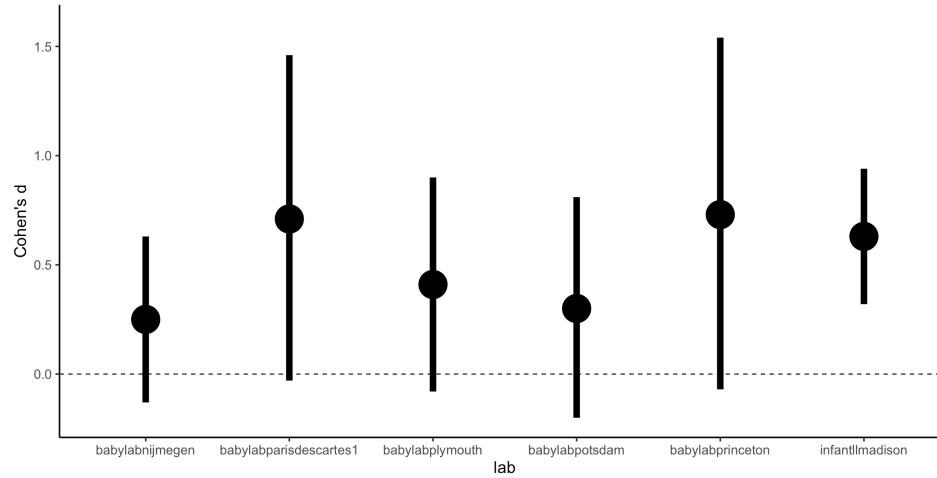
Midterm 9b



Many estimates of the size of an effect across many repeated experiments.



How do we summarize this pattern?



"The Madison Lab replicated the finding that infants prefer infant directed speech, while the other five labs did not."

That throws out a lot of information!!

Summarizing literatures is a more general challenge in psychology



consequently, it is underspecified in the mental lexicon. This predicts perceptual asymmetries such that labial mispronunciations of coronals (e.g., [bɔl] for /dɔl/) do not produce a mismatch ([bɔl] is accepted as an instance of /dɔl/), but coronal mispronunciations of labials do ([dɔl] is not accepted for /bɔl/). The results of numerous perceptual experiments are consistent with this prediction: labial mispronunciations prime coronal target words, but not vice versa, in cross-modal priming (Lahiri & Reetz, 2002). Similarly, event-related potential (ERP) studies have shown smaller ERPs to labial mispronunciations of coronals than vice versa (e.g., Cornell, Lahiri, & Eulitz, 2013).

Other work fails to support the predictions of FUL. Bonte, Mitterer, Zellagui, Poelmans, and Blomert (2005) reported smaller ERPs in response to a coronal-to-labial change compared to the opposite direction, but only when the non-words containing labials had a higher phonotactic probability than those containing coronals. With opposite phonotactic probabilities, this asymmetry reversed. In a series of three eye-tracking experiments, Mitterer (2011) found no evidence for asymmetric perception consistent with FUL, while a fourth experiment found an asymmetry predicted by phonotactic probability, but not FUL (but see Cornell et al., 2013). Based on this, Mitterer (2011) sug-

(Tsuji, et al. 2014)

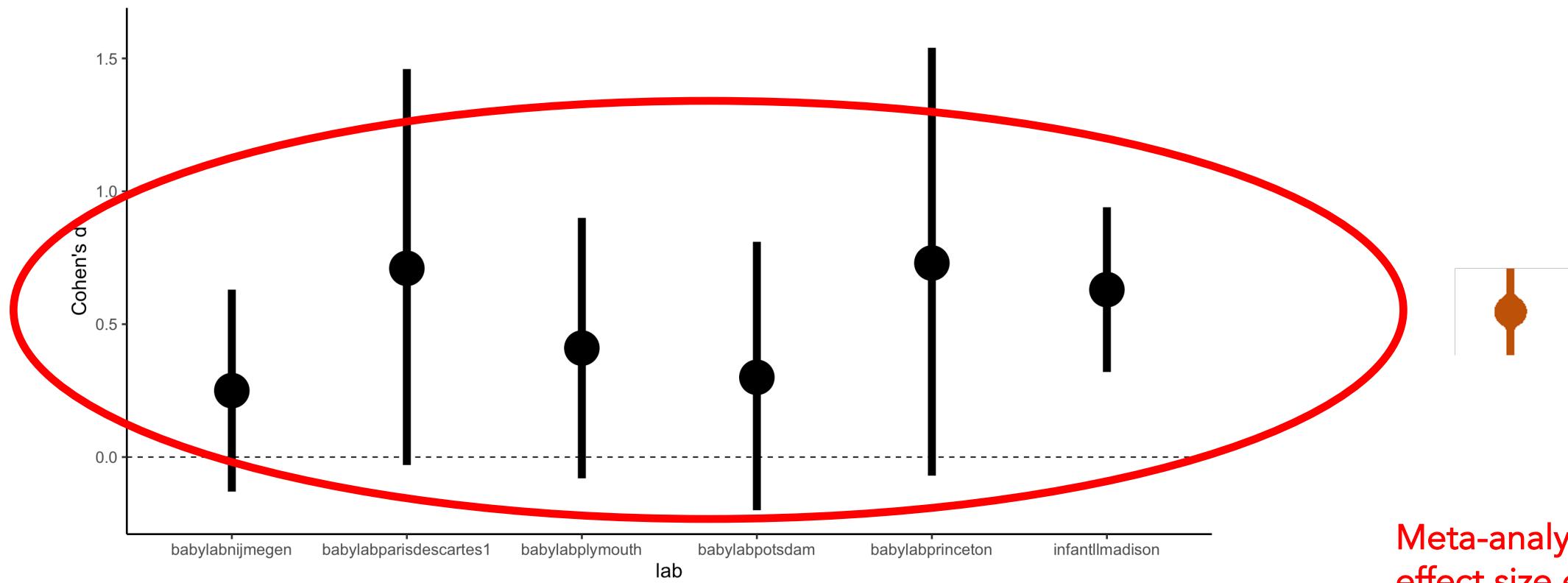
Psychological literatures
are almost always
conflicting

Qualitative literature
reviews are:

- not very precise
- difficult when there are many studies

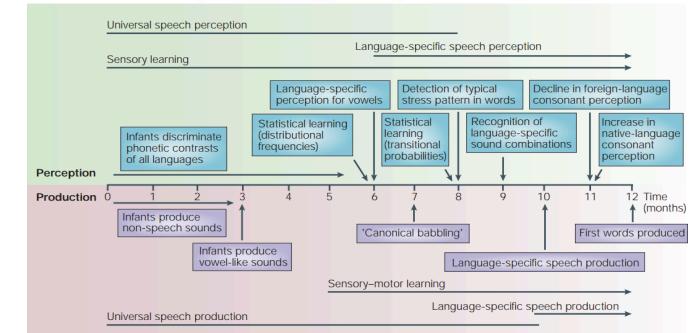
Meta-analysis

A quantitative approach to summarizing results across studies



Why do a meta-analysis?

1. Summarize what has been done in literature
2. Theory development – compare strength of different effects and moderating factors
3. Evaluate bias in literature (e.g. file drawer)
4. Estimate an effect size so you can determine a sample



How can we increase replicability?

Solution

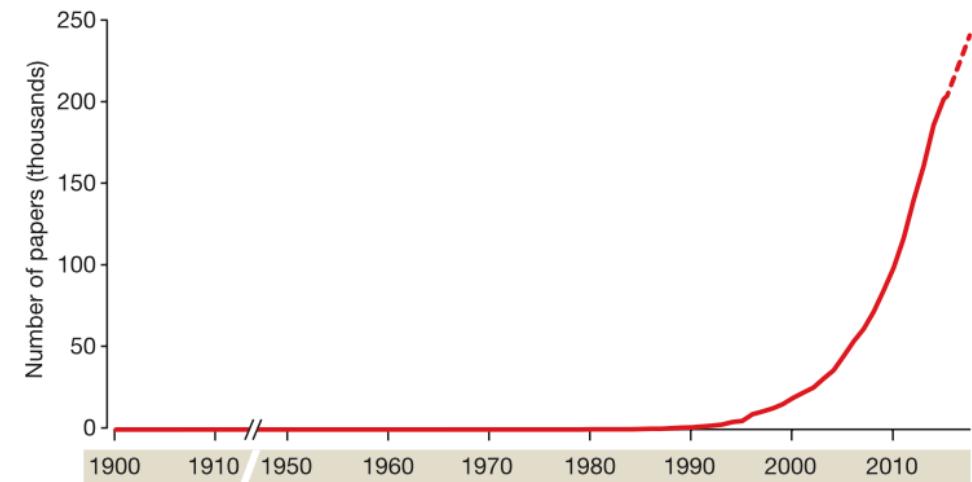
1. Data fraud
2. Analysis/reporting errors
3. Change in population effect size
4. Hidden moderators
5. File drawer
6. Data-dependent analysis (*p*-hacking)

Reproducibility practices
Strategies for reducing rates for failed replicates due to reasons that increase Type I error

N = ?

History of meta-analysis

- Mid-1970s: many studies had accumulated that were important to social decision policies
 - e.g. do students learn more when class sizes are smaller?
 - Research findings were conflicting, implications unclear, so test moderator variables, answers still unclear -> difficult to get funding
- Glass (1976): Research findings were not as conflicting as appeared
 - Using meta-analysis, reveals cumulative patterns
- The first “big data”



(Gurevitch, et al. 2018)

How are meta-analyses presented?

1. Stand-alone publications (~ literature review)
2. As part of an empirical paper (meta-analysis + new experiments)
3. Within-paper meta-analysis ("mini-meta-analysis")

Child Development, May/June 2001, Volume 72, Number 3, Pages 655–684

Meta-Analysis of Theory-of-Mind Development: The Truth about False Belief

Henry M. Wellman, David Cross, and Julianne Watson



Cognition
Volume 198, May 2020, 104191



PAPER

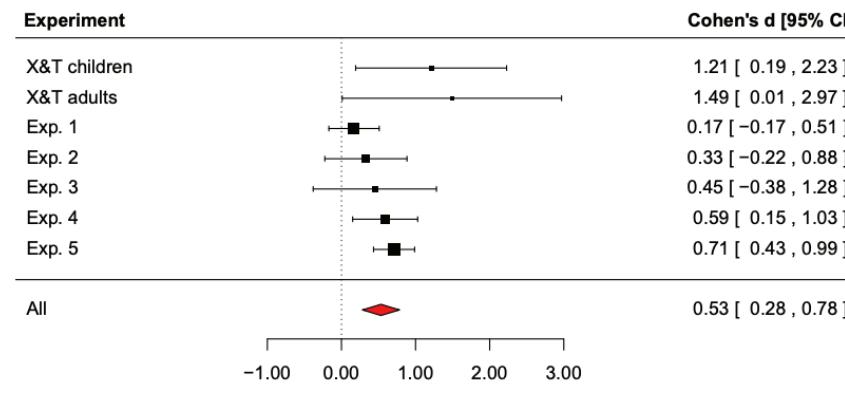
WILEY Deve

Original Articles

The profile of abstract rule learning in infancy: Me and experimental evidence

Hugh Rabagliati¹ | Brock Ferguson² | Casey Lew-Williams³

Molly Lewis ^a✉, Veronica Cristiano ^b, Brenden M. Lake ^{c, d}, Tammy Kwan ^{c, d}, Michael C. Frank ^e



(Lewis & Frank, 2016)

How do you do a meta-analysis?

(1a) Collect studies

In our first experiment, 24 8-month-old infants from an American-English language environment were familiarized with 2 min of a continuous speech stream consisting of four three-syllable nonsense words (hereafter, "words") repeated in random order (16). The speech stream was generated by a speech synthesizer that simulated adult voice at a rate of 270 syllables per minute (180 words in total). The synthesizer provided no acoustic information about word boundaries, resulting in a continuous stream of concatenated consonant-vowel syllables with no pauses, stress differences, or any other acoustic or prosodic cues to word boundaries. A sample of the speech stream is the orthographic string *bikakapadatogka-bubububu...*. The only cues to word boundaries were the transitions between different two-syllable pairs, which were¹ assigned to one of two counterbalanced language conditions: Language 1A within words (1.1 in all cases for each *bikakapadatogka*) and Language 1B (between words 0.3 in all).

To aid learning, each infant was presented with a sequence of one of four syllable strings on each test trial. Of these three-syllable strings were² from the artificial language presenting familiarity, and two were the labile "nonwords" that contained the syllables of the target words familiarized in the order in which they appear words (17).

The infants showed a significant trial discrimination between word and nonword times.

Table 1. Mean time spent listening to the nonwords and experiment 2 (words versus times).

Experiment	Mean listening times
1	7.97 (SE = 0.41)
2	6.77 (SE = 0.44)

We created two counterbalanced languages to control for arbitrary listening preferences at test. Language 1A consisted of three identical blocks of 12 grammatically correct and semantically meaningful standard Italian sentences (see the Appendix for sentence lists). These sentences contained the words *fuga* and *melo*, which both occurred six times in each block of 12 sentences. The component syllables of *fuga* and *melo* never appeared without each other (i.e., *fu* never appeared in the absence of *ga*, and vice versa).

Recall that the TP of, for example, *fuga* corresponds to:

$$TP(fuga|fu) = \frac{f(fuga)}{f(fu)}.$$

Because *fu* never appeared without *ga*, the internal TP of *fuga* (and of *melo*) was 1.0. Two other words, *pane* and *tema*, and their component syllables, were never presented in the Language 1A familiarization passages (TP = 0). In the counterbalanced Language 1B, *pane* and *tema* each occurred six times per block (TP = 1.0), while *fuga* and *melo* (and their component syllables) never occurred (TP = 0). This design is thus exactly analogous to the original Jusczyk and Aslin (1995) study.

(1b) Code

	A	B	C	D	E
1	study_ID	long_cite	short_cite	same_infant	coder
2	SaffranAslinNewport1996	Saffran, J. R., Aslin, R. N., & Nevin Saffran, Aslin, & Newport (1996)			Reference
3	SaffranAslinNewport1996	Saffran, J. R., Aslin, R. N., & Nevin Saffran, Aslin, & Newport (1996)			Reference
4	PelucchiHaySaffran2009a	Pelucchi, B., Hay, J. F., & Saffran Pelucchi, Hay, & Saffran (2009)			Reference
5	PelucchiHaySaffran2009a	Pelucchi, B., Hay, J. F., & Saffran Pelucchi, Hay, & Saffran (2009)			Reference
6	PelucchiHaySaffran2009a	Pelucchi, B., Hay, J. F., & Saffran Pelucchi, Hay, & Saffran (2009)			Reference
7	PelucchiHaySaffran2009b	Pelucchi, B., Hay, J. F., & Saffran Pelucchi, Hay, & Saffran (2009b)			Reference

(2) Aggregate

$$ES = .6$$

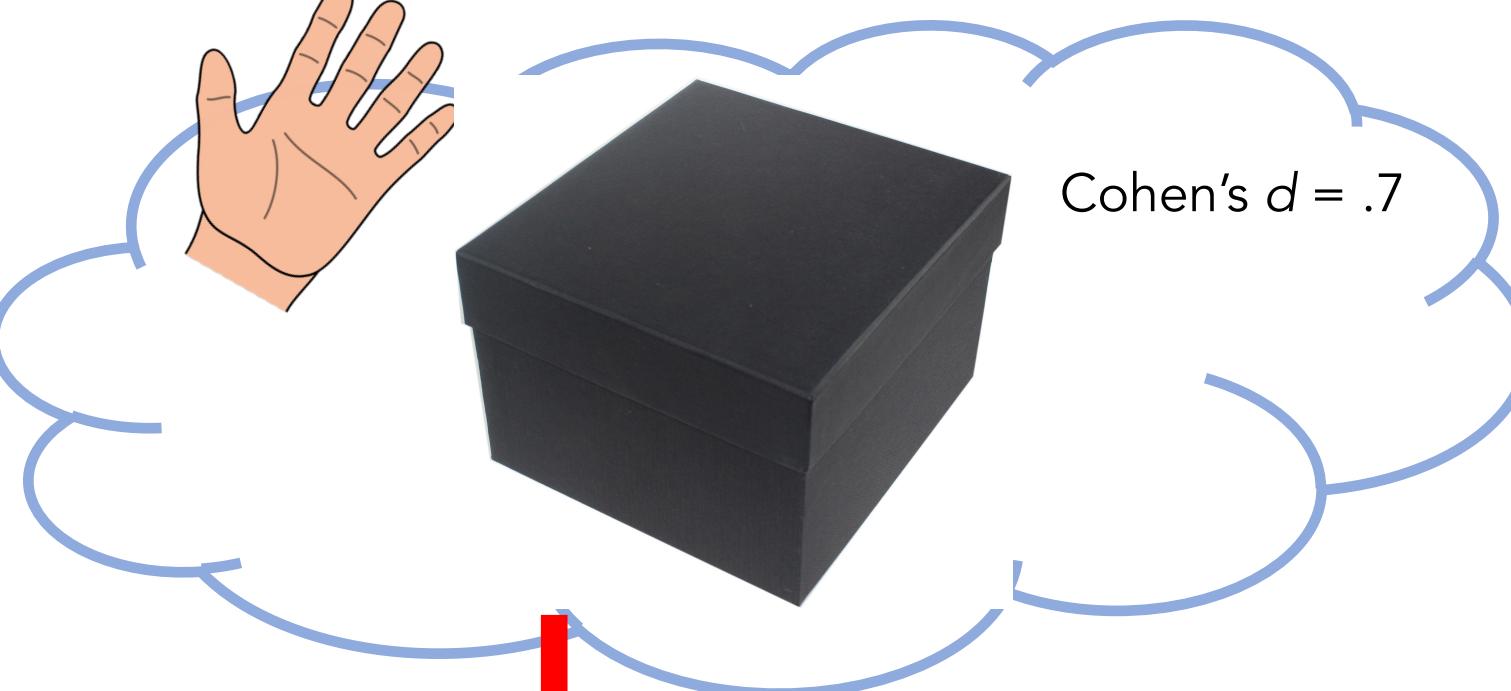
How do you aggregate?

- The goal of a meta-analysis is to estimate the true population effect size
- Treat each study as an sample effect size from a population of studies
- Aggregate using quantitative methods (e.g. averaging)
- Get point estimate of the true effect size with measure of certainty

More precise estimate of effect size than from single study.

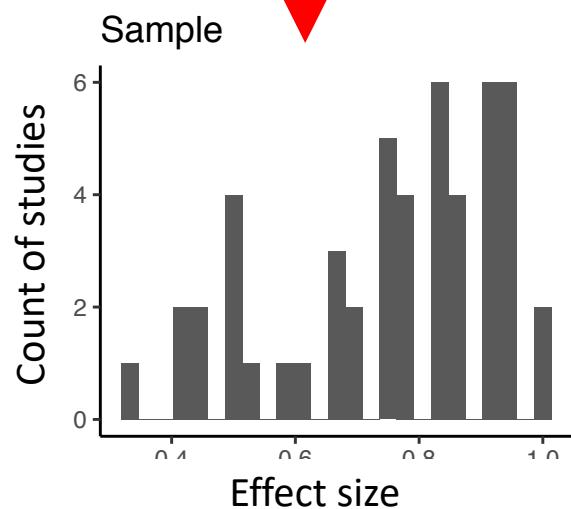
Population

All the studies we could have run



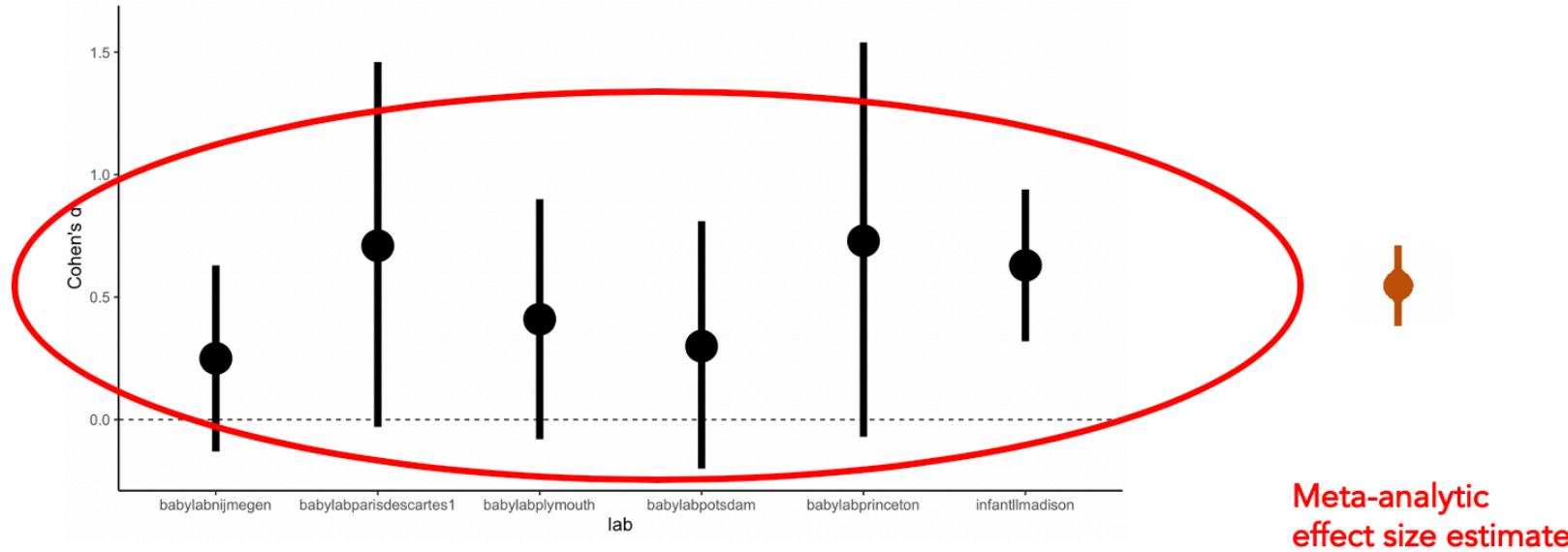
Sample

All the studies we did run (i.e. the ones in the literature)



Use samples to estimate population.

How do we go from samples to an estimate of the population?



- Basically, just average all the effect sizes in our sample.
- Weighting by sample size

An example meta-analysis: Mutual exclusivity

Cognition 126 (2013) 39–53

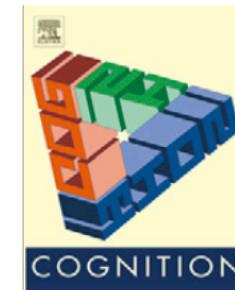


ELSEVIER

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT



Fast mapping, slow learning: Disambiguation of novel word–object mappings in relation to vocabulary learning at 18, 24, and 30 months

Ricardo A.H. Bion ^{a,*}, Arielle Borovsky ^{a,b}, Anne Fernald ^a

^a Stanford University – Department of Psychology, 450 Serra Mall, Stanford, CA 94305, United States

^b University of California, San Diego – Center for Research in Language, 9500 Gilman Drive MC 0526, La Jolla, CA 92093-0526, United States

An example: Mutual exclusivity

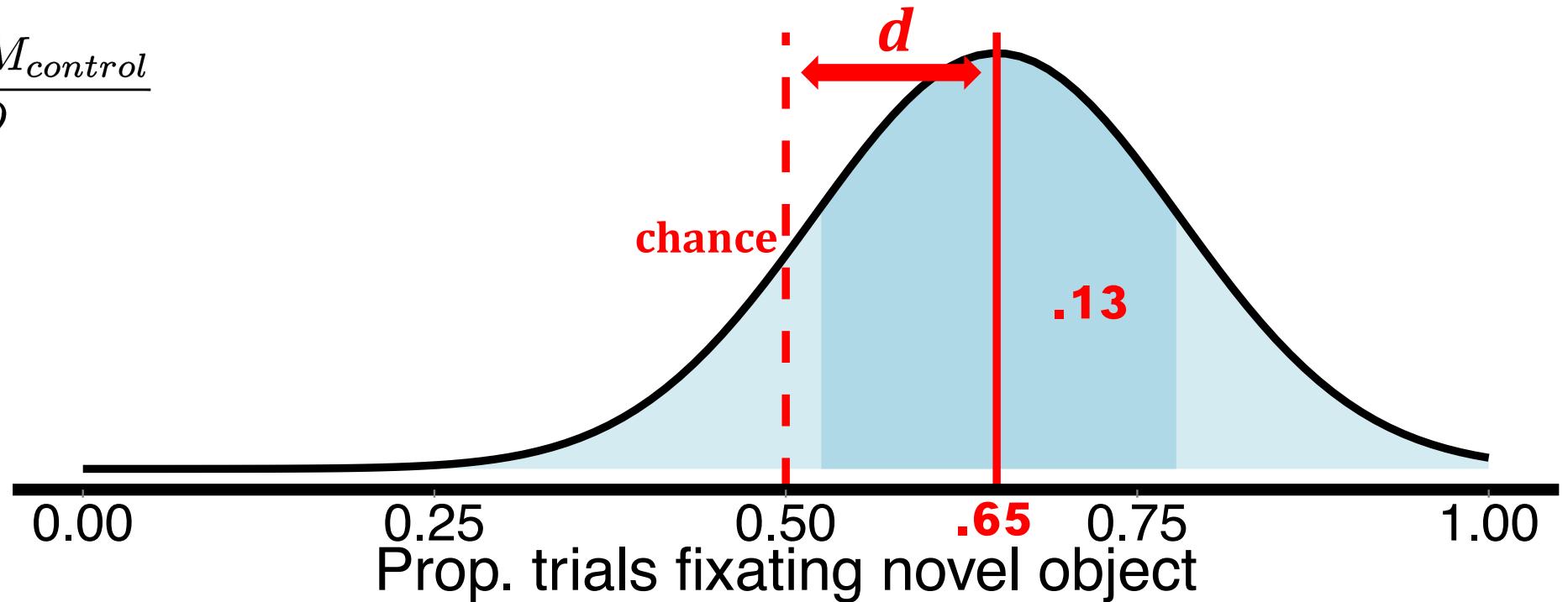
Where's the dofa?



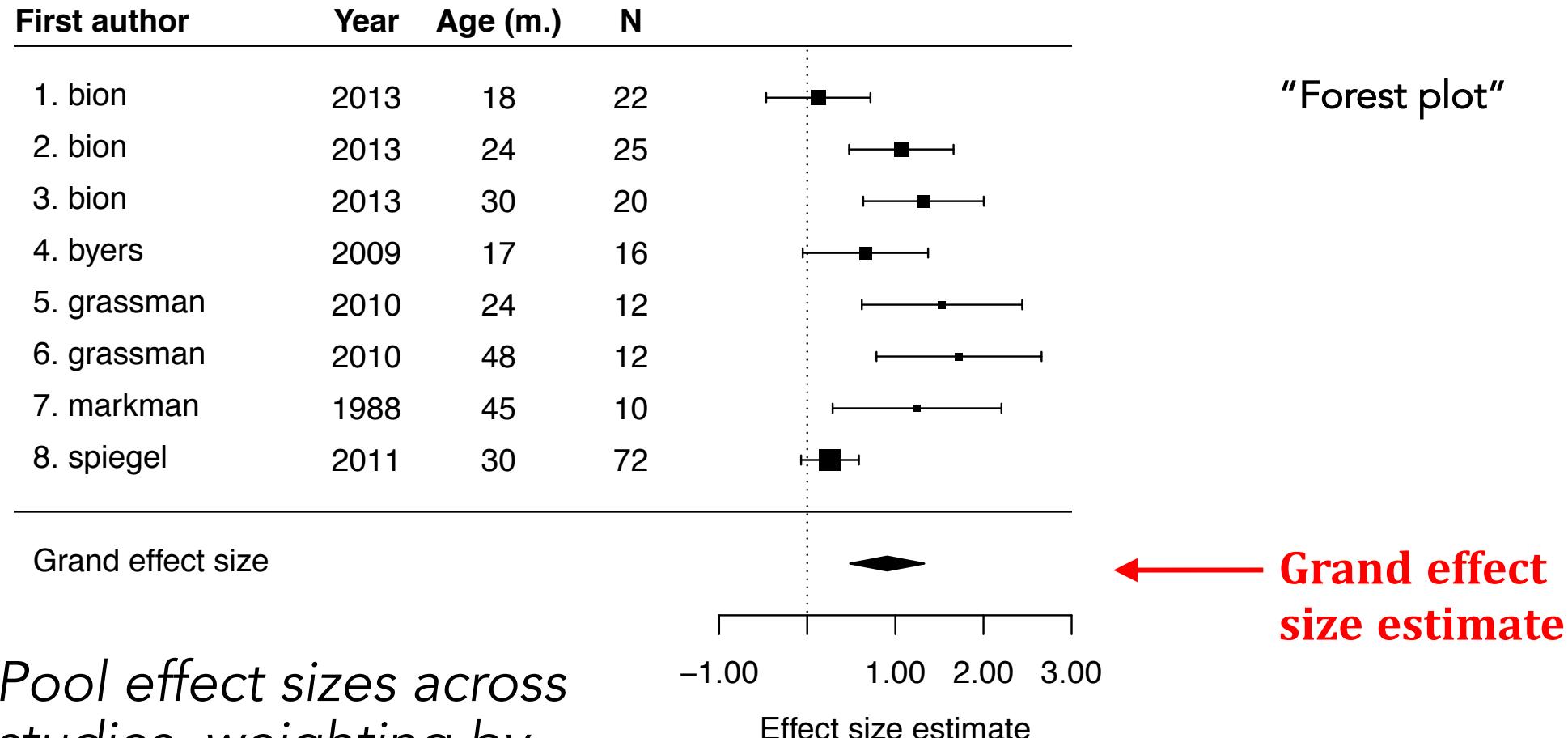
Bion, et al. (2013)

For 24 mo, mean proportion of trials fixating on novel object = .65 (SD = .13)

$$d = \frac{M_{exp} - M_{control}}{SD}$$



Mutual exclusivity meta-analysis



More practice coding effect sizes from papers

J. Child Lang. 28 (2001), 787–804. © 2001 Cambridge University Press
DOI: 10.1017/S0305000901004858 Printed in the United Kingdom

NOTE

By any other name: when will preschoolers produce several labels for a referent?*

GEDEON O. DEÁK AND LOULEE YEN

Vanderbilt University

JEREMY PETTIT

David Lipscomb University

(Received 23 February 2000. Revised 8 December 2000)

Your job #2

- In your breakout groups, calculate an effect size for the two age groups in Deak, et al. (2001) in Experiment 1.
- If you have time, you can also calculate the confidence intervals on the effect sizes.

Paper: <https://bit.ly/3a8GmuH> (also linked on website)

Next Time: Introduction to conducting your own meta-analysis

Wrap up coding of Deak et al. paper if you have not

Reading: (4 pages)

1

RESEARCH SYNTHESIS AS A SCIENTIFIC PROCESS

HARRIS COOPER
Duke University

LARRY V. HEDGES
Northwestern University