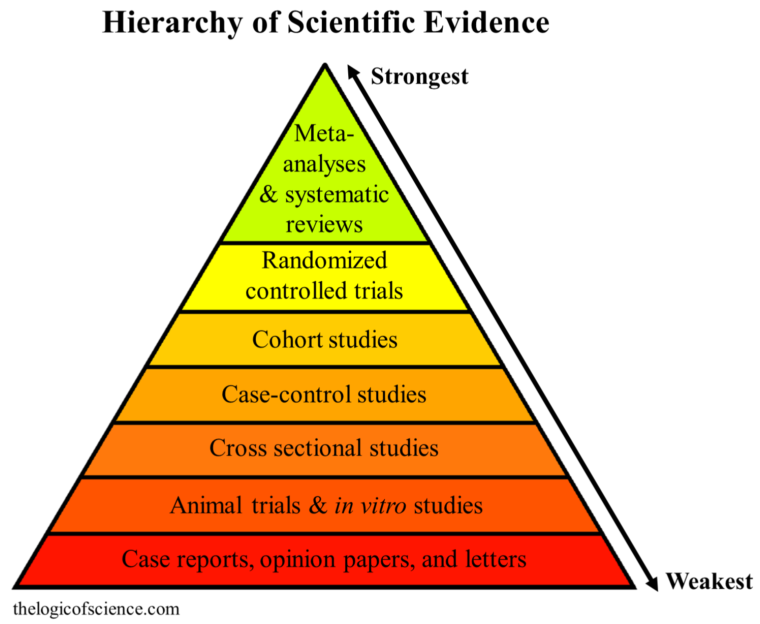
Introduction to Meta-Analysis

25 October 2021

Modern Research Methods



Overview of course

1) **The Process of Cumulative Science**

2) **The Single Experiment** – Experimental data, 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

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 mostly compared outcomes for two replications – but what if we have <u>many replications</u>?

Overview of course

1) **The Process of Cumulative Science**

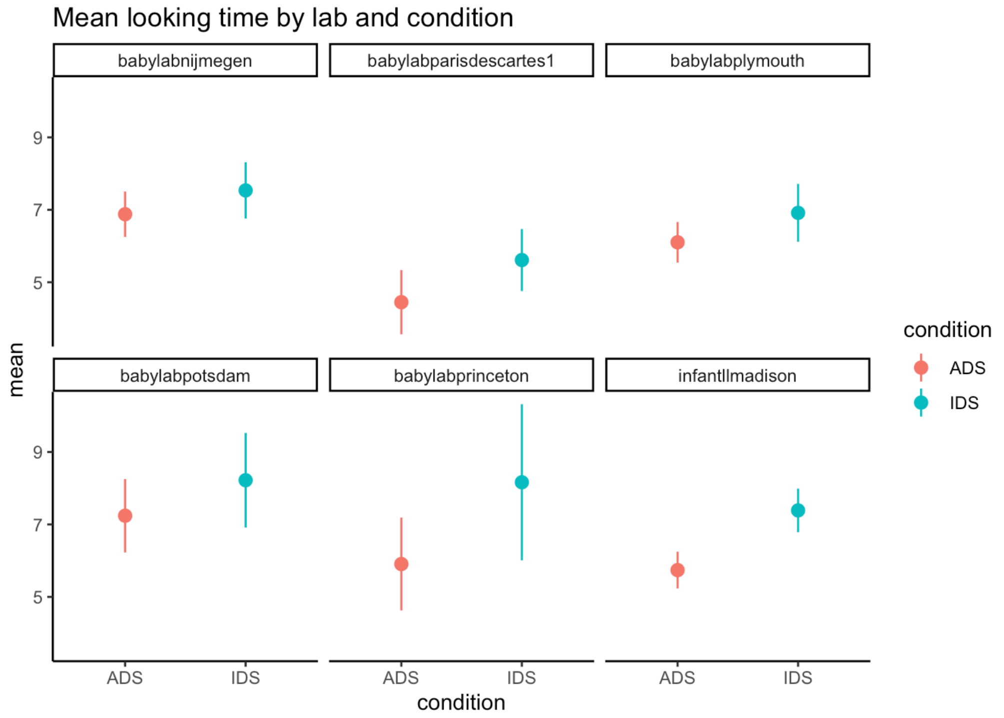
2) **The Single Experiment** – Experimental data, tools in R for working with data and plotting data, reproducibility

3) **Repeating an Experiment** – Intro to statistical concepts, replication of experiments

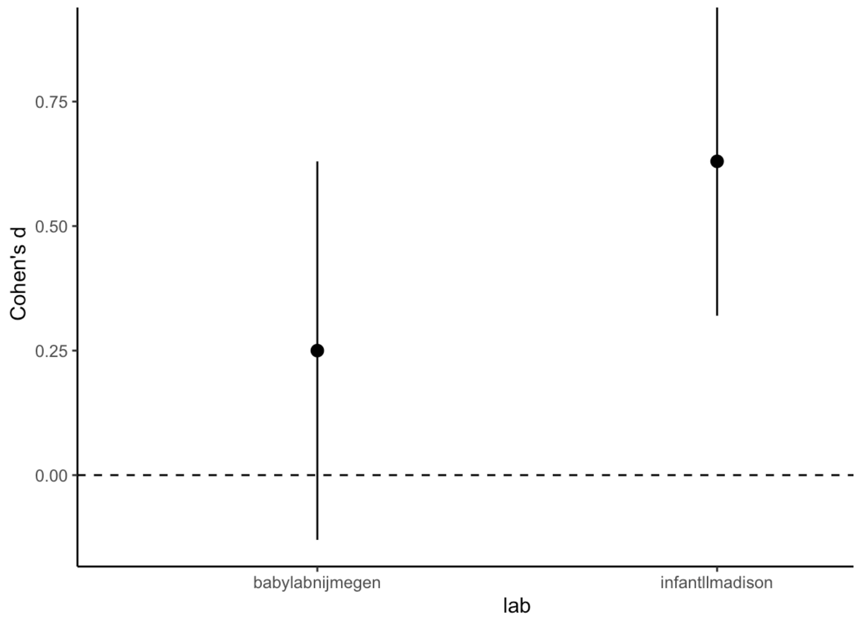
4) **Aggregating Many Experiments** – Meta-analysis

A repeated experiment

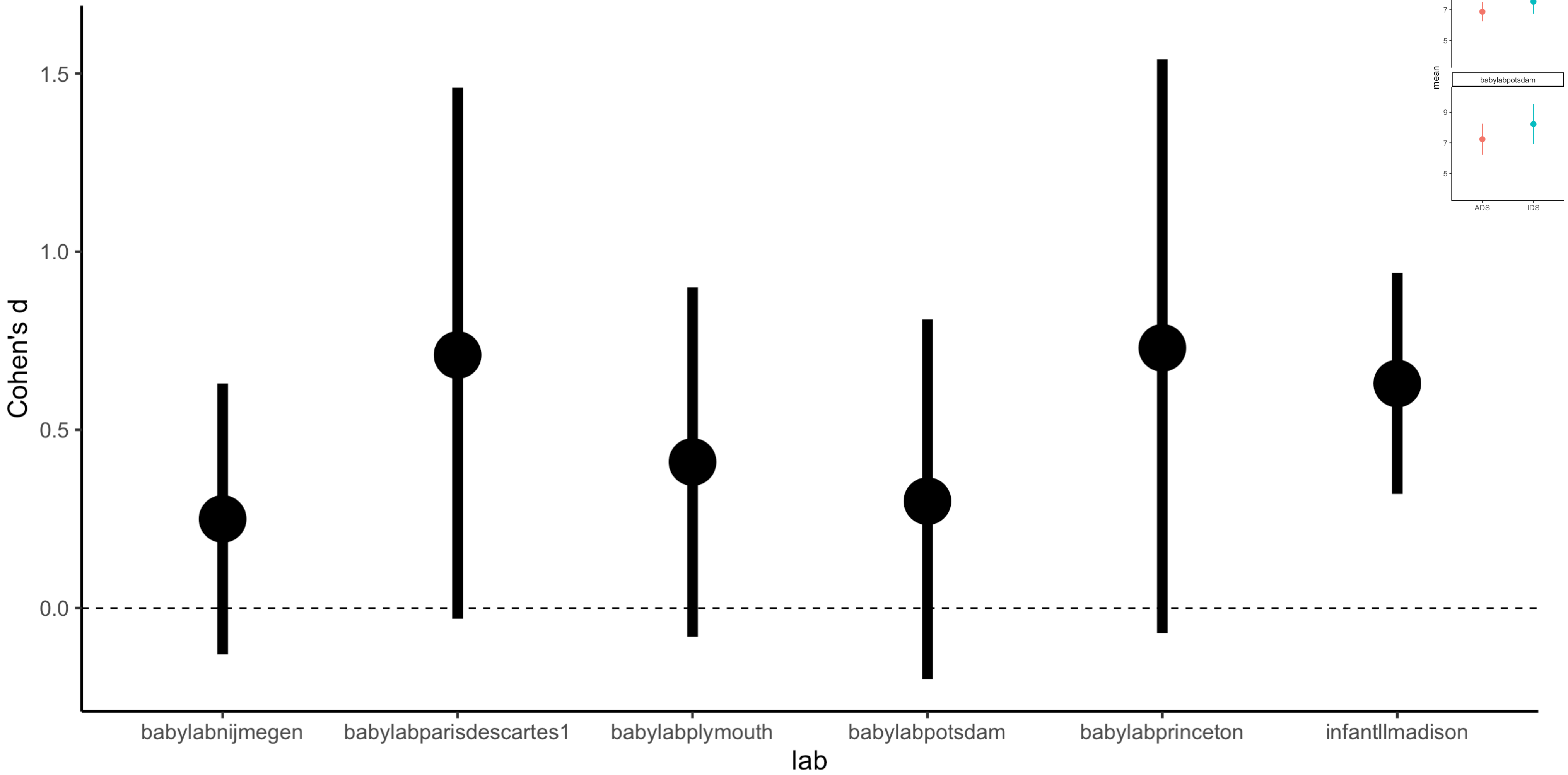
Assignment 5

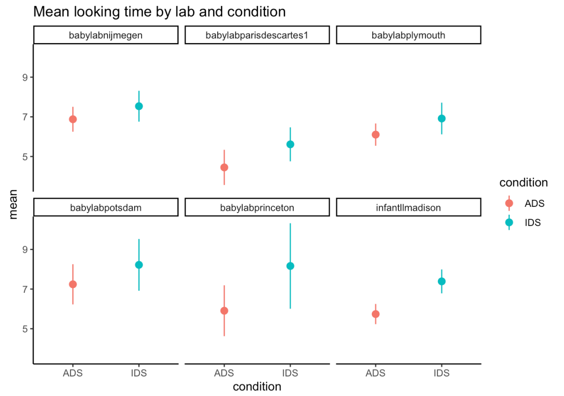


Midterm 11b

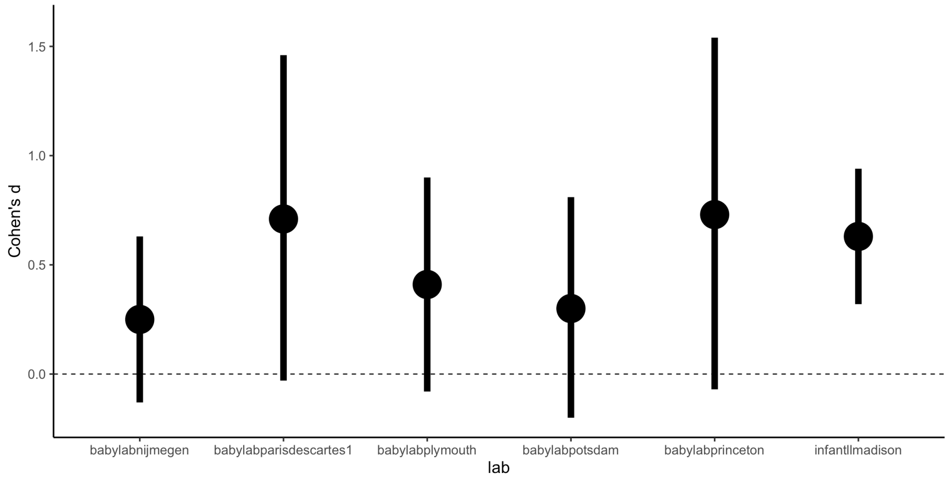


Many estimates of the size of an effect across many repeated experiments.





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)

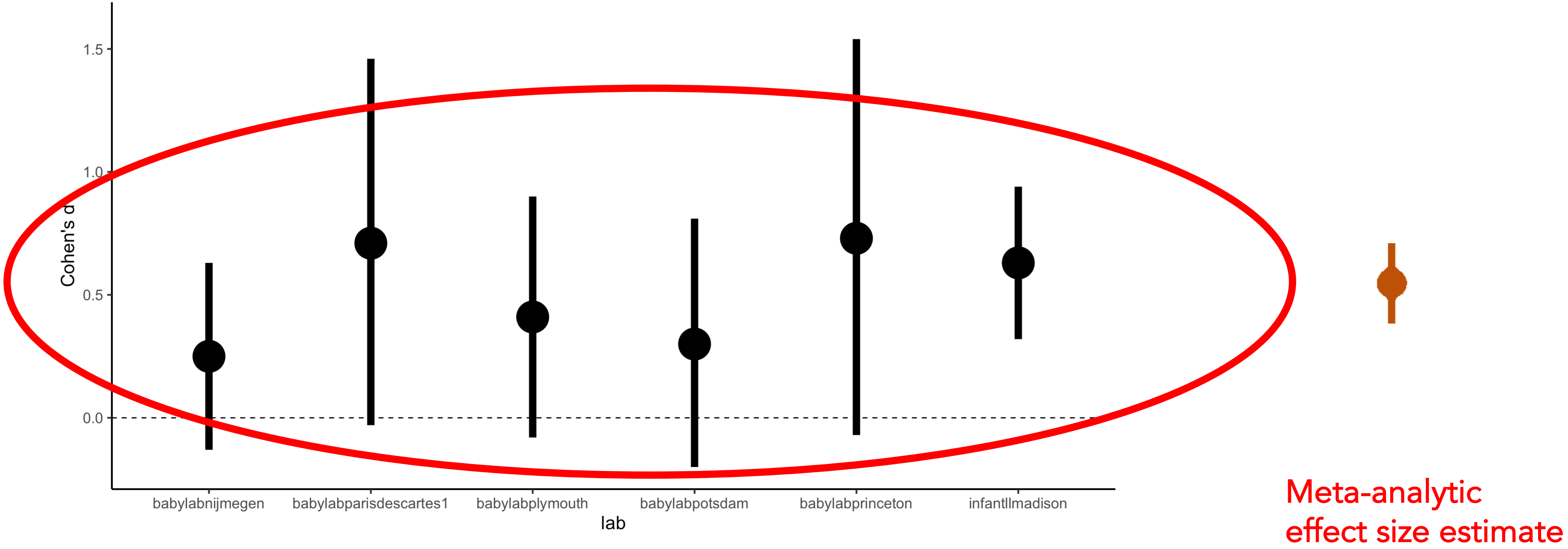
Psychological literatures are almost always conflicting

Qualitative literature reviews are:

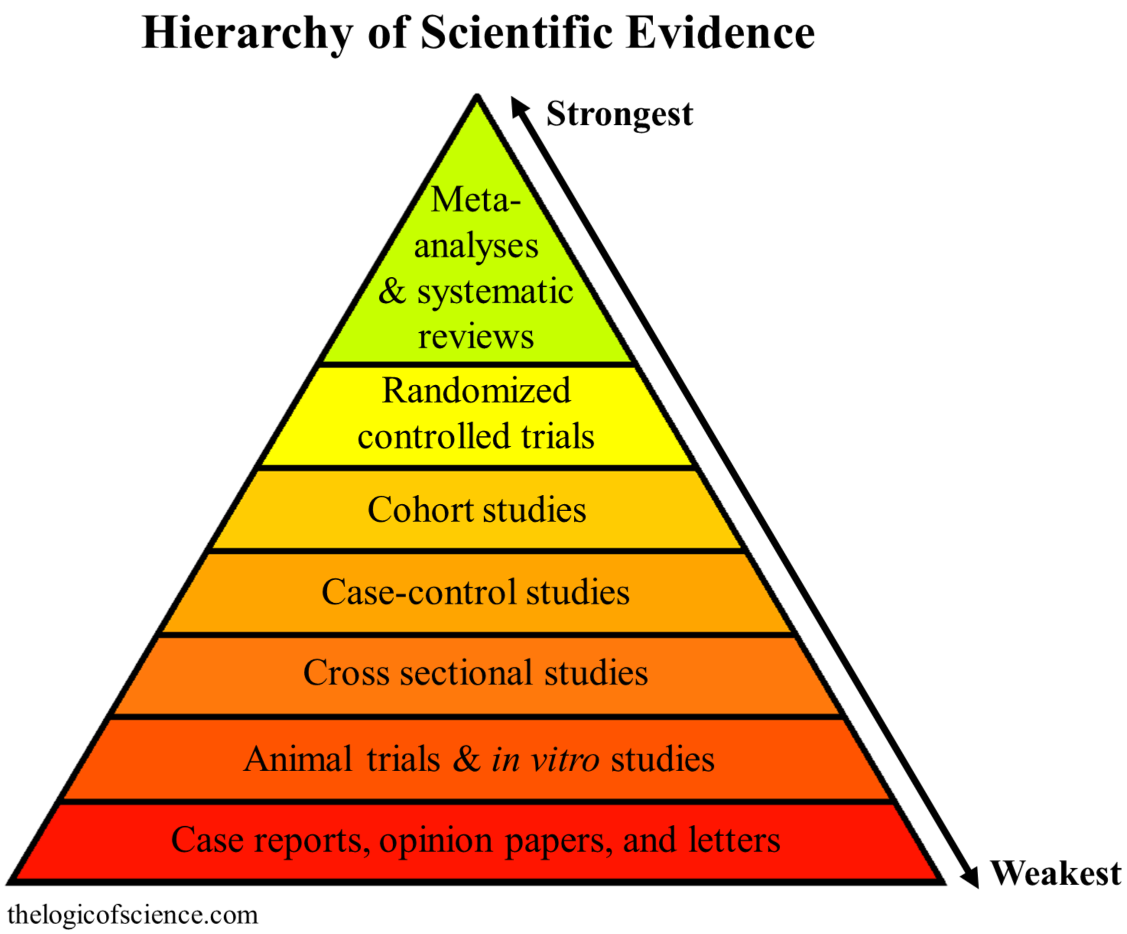
- not very precise
- difficult when there are many studies

Meta-analysis

A quantitative approach to summarizing results across studies



Meta-analysis at the top of the evidence hierarchy



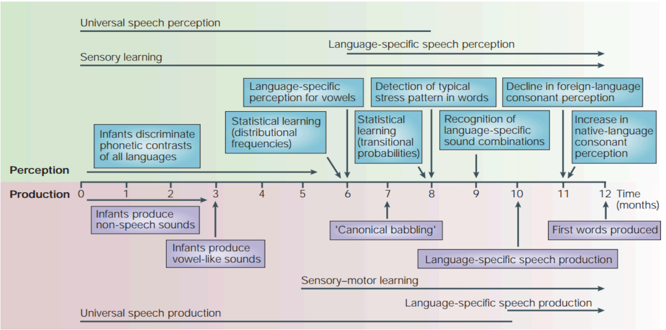
Why do a meta-analysis?

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

Reproducibility practices

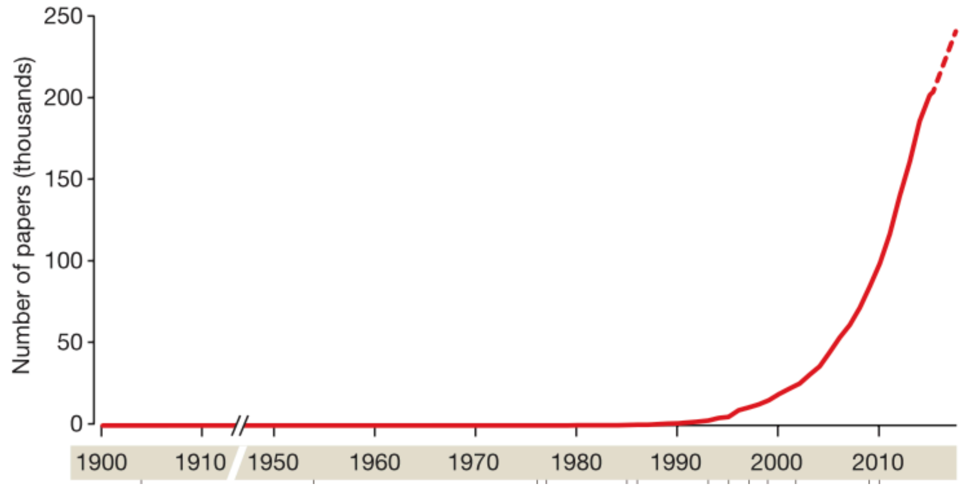
Strategies for reducing rates for failed replicates due to reasons that increase Type I error

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)

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 experimental 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 Julanne Watson

PAPER

WILEY

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

Hugh Rabagliati[1] | Brock Ferguson[2] | Casey Lew-Williams[3]

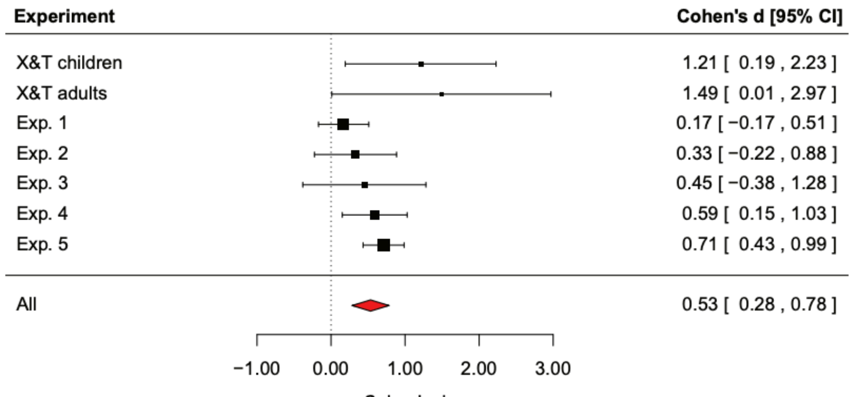
ELSEVIER

Cognition
Volume 198, May 2020, 104191

Original Articles

The role of developmental change and linguistic experience in the mutual exclusivity effect ☆

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 in a monotone female 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 coarticulated 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 *bidakupadotigolabubidaku*. . . . The only cues to word boundaries were the transitional probabiliti tween syllable pairs, which were within words (1.0 in all cases, for exa *bida*) than between words (0.33 in all for example, *kupa*).

To assess learning, each infant wa sented with repetitions of one of four syllable strings on each test trial. T these three-syllable strings were "v from the artificial language presente ing familiarization, and two were thr lable "nonwords" that contained the syllables heard during familiarizatio not in the order in which they appea words (*17*).

The infants showed a significan trial discrimination between word an

rent syllable sequences) from syllable strings spanning word boundaries (that is, syllable sequences occurring more rarely). To take an English example, *pretty#baby*, we wanted to see if infants can distinguish a word-internal syllable pair like *pretty* from a word-external syllable pair like *ty#ba*.

Another 24 8-month-old infants from an American-English language environment were familiarized with 2 min of a continuous speech stream consisting of three-syllable nonsense words similar in structure to the artificial language used in our first experiment (*19*). This time, however, the test items for each infant consisted of two words and two "part-words." The part-words were created by joining the final syllable of a word to the first two syllables of

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

Experiment	Mean liste
	Familiar items
1	7.97 (SE = 0.41)
2	6.77 (SE = 0.44)

Participants were assigned to one of two counterbalanced language conditions: Language 1A and Language 1B. Eighteen additional infants were tested and excluded for the following reasons: fussiness (14), experimental error (3), and not paying attention (1). Two additional infants showed looking time preferences > 3 *SD* from the mean (one in each language group with preferences in opposite directions), and were excluded from the analyses.

Apparatus and stimulus materials—Four Italian words with a strong–weak stress pattern were selected for use in this study: *fuga, melo, pane*, and *tema* (see Table 1). Although these words were phonetically legal in English, the passages in which they were presented contained non-English phonetic features (e.g., a trill, a voiced alveolar affricate, and a palatal nasal).

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:

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

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 each 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., & Nev	Saffran, Aslin, & Newport (1996)		Reference
3	SaffranAslinNewport1996	Saffran, J. R., Aslin, R. N., & Nev	Saffran, Aslin, & Newport (1996)		Reference
4	PelucchiHaySaffran2009a	Pelucchi, B., Hay, J. F., & Saffrar	Pelucchi, Hay, & Saffran (2009)		Reference
5	PelucchiHaySaffran2009a	Pelucchi, B., Hay, J. F., & Saffrar	Pelucchi, Hay, & Saffran (2009)		Reference
6	PelucchiHaySaffran2009a	Pelucchi, B., Hay, J. F., & Saffrar	Pelucchi, Hay, & Saffran (2009)		Reference
7	PellucchiHaySaffran2009b	Pelucchi, B., Hay, J. F., & Saffrar	Pelucchi, Hay, & Saffran (2009b)		Reference

(2) Aggregate

ES = .6

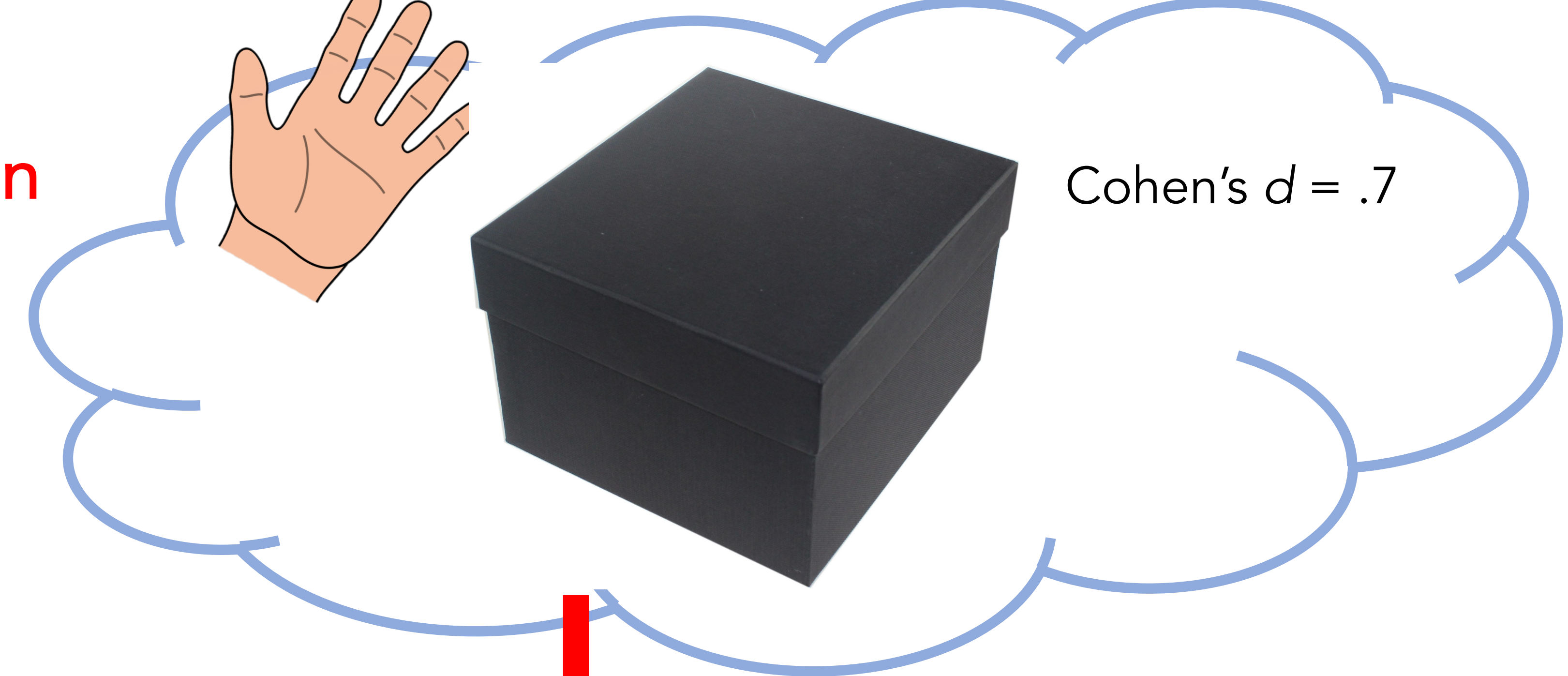
How do you aggregate?

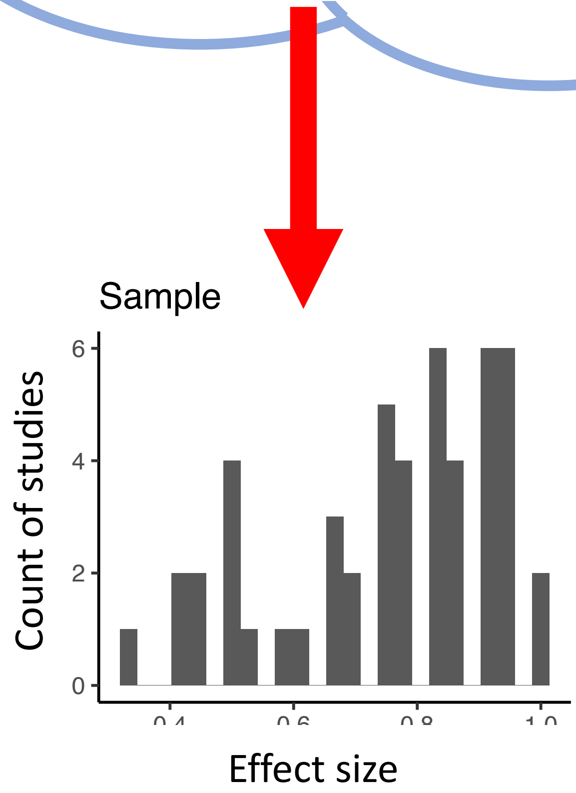
- The goal of a meta-analysis is to estimate the true population effect size
- Treat each study as an sample effect size from a population of studies
- Aggregate using quantitative methods (e.g. averaging)
- Get point estimate of the true effect size with measure of certainty

More precise estimate of effect size than from single study.

Population

All the studies we could have run



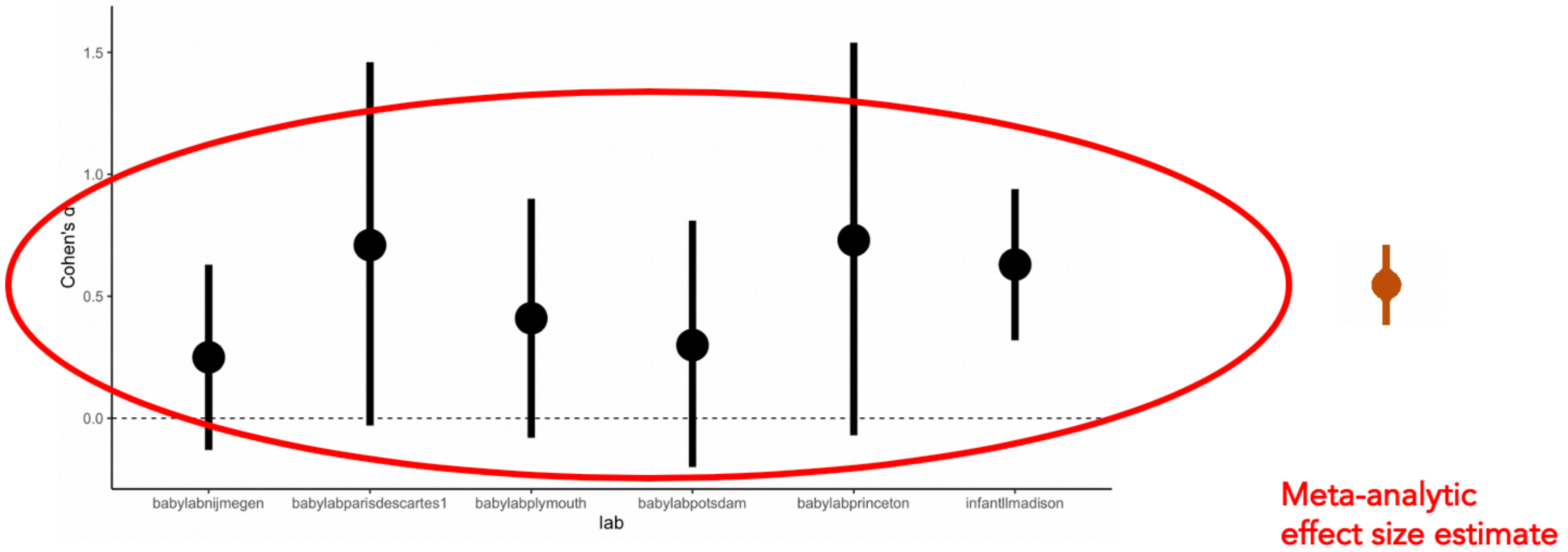


Sample

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

What to aggregate in the meta-analysis?

- P-values give you a yes/no answer – is the difference significant or not? (“vote counting”)
- Effect sizes (e.g. Cohen’s *d)* – how big is the effect and what direction is it in?

- “Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude – not just, does a treatment affect people, but how much does it affect them.” - Gene Glass

Review: Effect size measures

- For any statistical test you conduct, can compute effect size (in principle)
- ES depends on design
- Can convert between ES metrics

Review: Cohen's *d*

Standardized measure of the size of an effect when you have a categorical IV and a continuous DV.

Cohen's *d*:

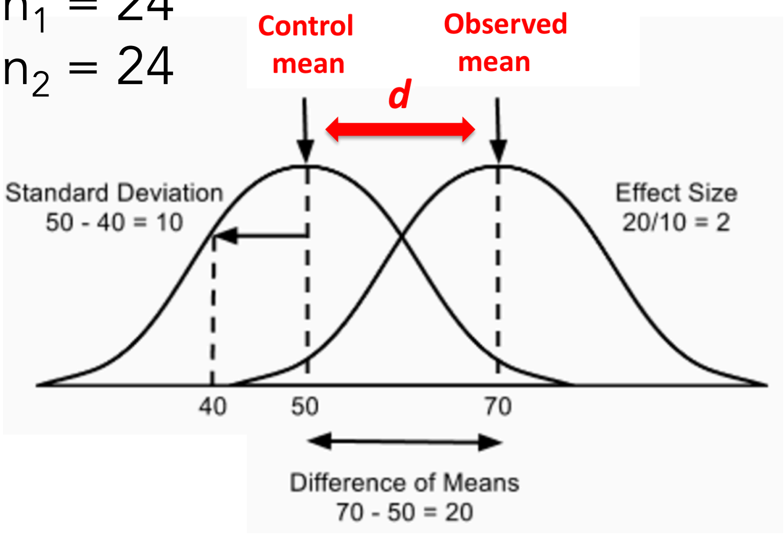
$$\text{Effect Size} = \frac{\textbf{diff. between means}}{\textbf{standard dev.}}$$

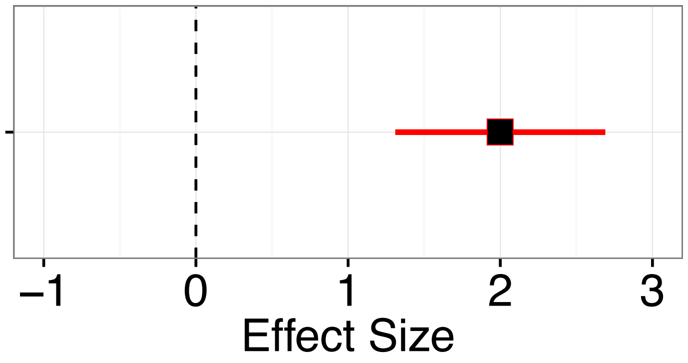
$$d = \frac{M_{group1} - M_{group2}}{SD_{pooled}}$$

$$SD_{pooled} = \sqrt{(SD^2_{group1} + SD^2_{group2})/2}$$

Cohen's *d* confidence interval

$n_1 = 24$
$n_2 = 24$





$$
\begin{aligned}
var_d &= \frac{n_1 + n_2}{n_1 * n_2} + \frac{d^2}{2(n_1 + n_2)} \\
&= \frac{24 + 24}{24 * 24} + \frac{2^2}{2(24 + 24)} \\
&= .125
\end{aligned}
$$

$$
\begin{aligned}
CI(d) &= Est(d) \pm z_{(\alpha/2)} * \sqrt{var(d)} \\
&= 2 \pm 1.96 * .35 \\
&= 2 \pm .69
\end{aligned}
$$

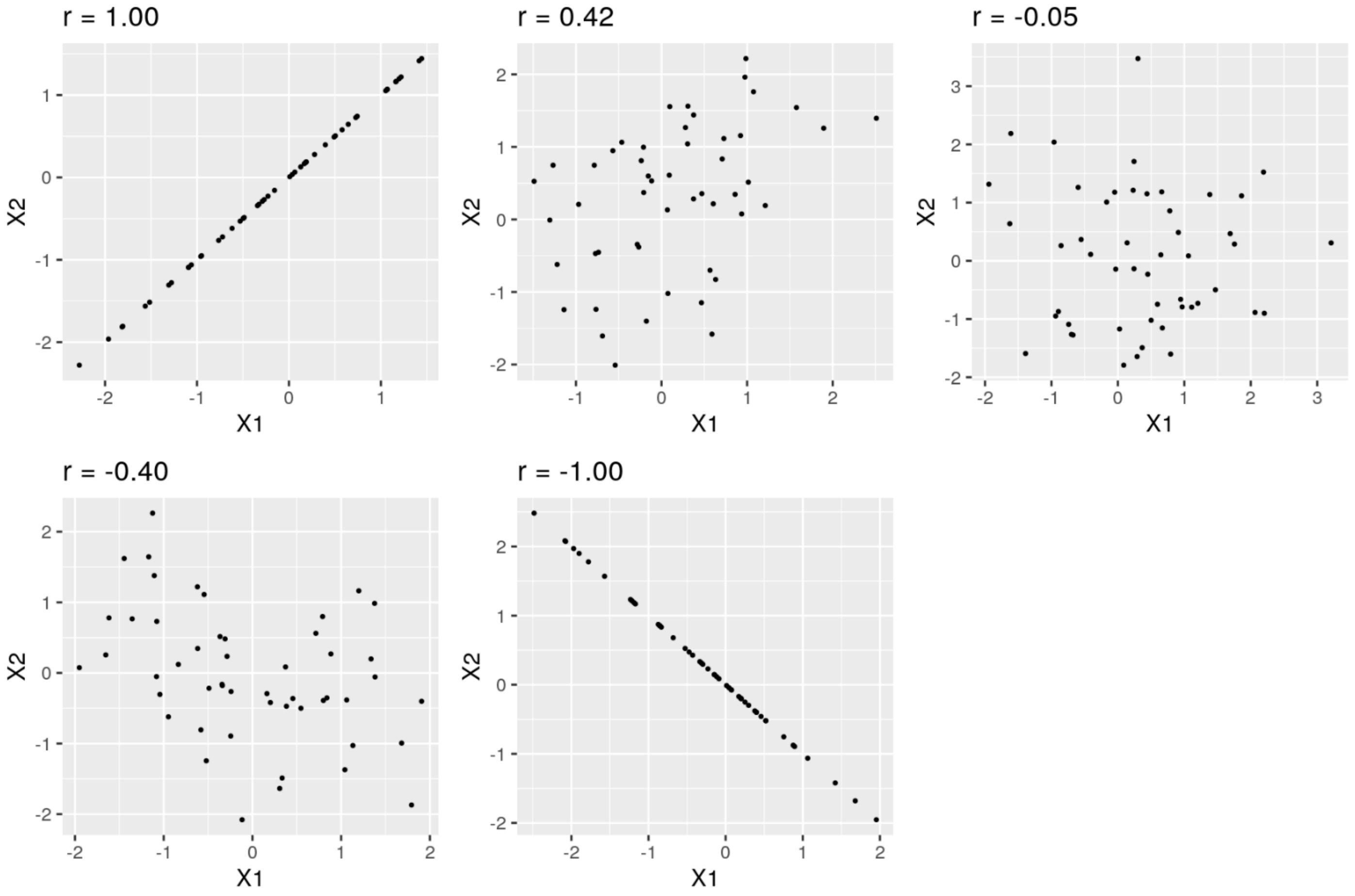
Pearson's *r*

Correlation coefficient

Standardized measure of th
size of an effect when you h
a continuous IV and a
continuous DV.

Ranges from -1 to 1

Typically don't have to calculate
it (reported in paper)



An example meta-analysis: Mutual exclusivity

Cognition 126 (2013) 39–53

Contents lists available at SciVerse ScienceDirect

ELSEVIER

Cognition

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

COGNITION

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

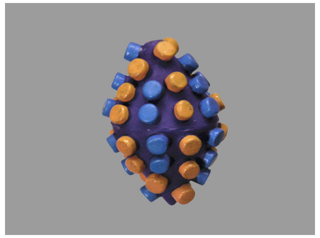
Ricardo A.H. Bion [a,*], Arielle Borovsky [a,b], Anne Fernald [a]

[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

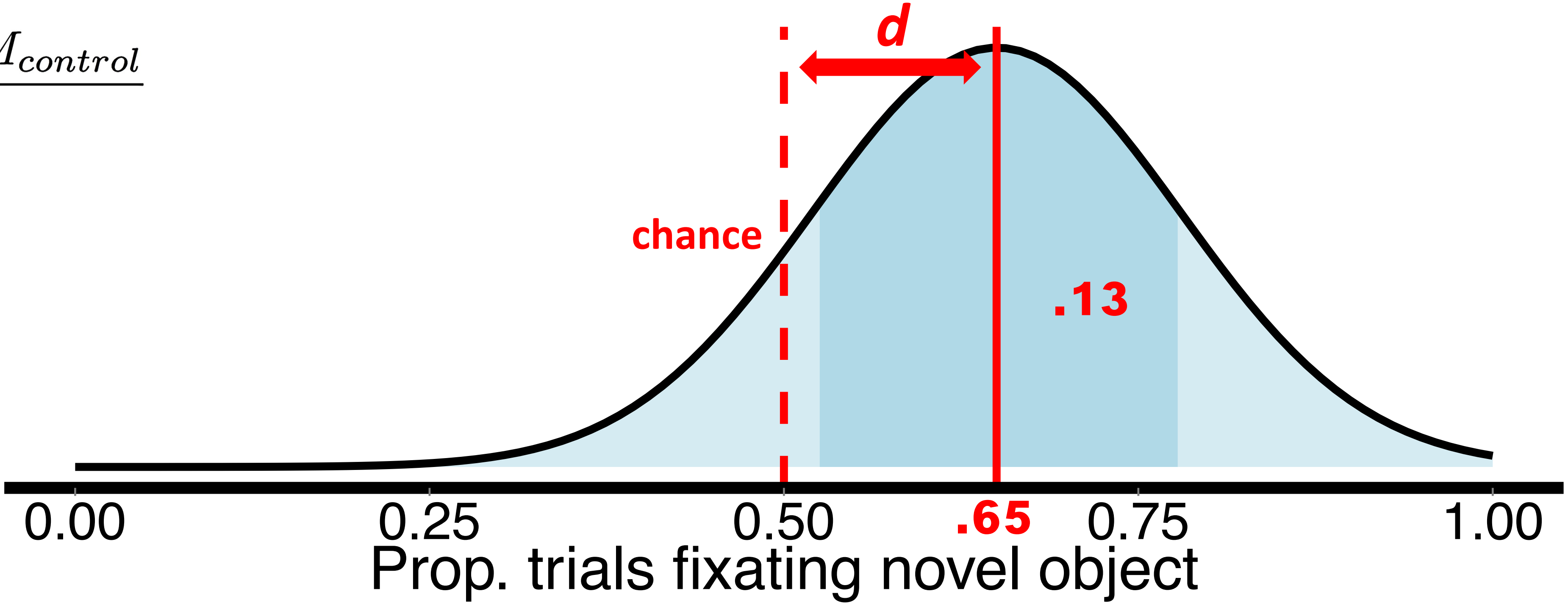
Where's the dofa?



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

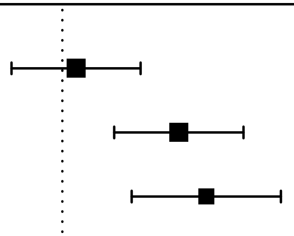
<u>Bion, et al. (2013)</u>

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



Mutual exclusivity meta-analysis

First author	Year	Age (m.)	N
1. bion	2013	18	22
2. bion	2013	24	25
3. bion	2013	30	20



"Forest plot"

Grand effect size estimate

More practice coding effect sizes from papers

J. Child Lang. **28** (2001), 787–804.
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)

Practice coding effect sizes

- In groups, calculate an effect size for the two age groups in Deak, et al. (2001) in Experiment 1.
- Note that another name for ”mutual exclusivity” is “disambiguation”
- You’ll have to dig into the paper a little bit to find the relevant numbers.
- If you have time, you can also calculate the confidence intervals on the effect sizes.

Next Time: Choosing an MA topic

- Guest lecture by Anjie Cao (former MRM student)!
- Read her (almost published!) paper

Quantifying the syntactic bootstrapping effect in verb learning: A meta-analytic synthesis

AUTHORS
Anjie Cao, Molly Lewis

(in press, *Developmental Science*)