Meta-Analysis, Open Science, and Publication Bias

Wolfgang Viechtbauer

Open Science Workshop:
Promoting transparency & replicability in research

University of Tübingen
September 22, 2022

Wolfgang Viechtbauer
Department of Psychiatry and Neuropsychology
Maastricht University, The Netherlands
https://www.wvbauer.com/

Goals of this Talk

- not to teach you meta-analysis (MA)
- provide a bit of motivation for MA itself
- make the point that MA was built on some open-science principles
- describe some critiques of MA (and responses)
- talk about publication bias
- give some recommendations for open-science practices in the context of MA

The Information Explosion

- 2,300 biomedical journals in 1940
- now there are close to 25,000
- 27,000+ RCTs per year
- similar growth in other disciplines
- rough estimates:
  - number of articles doubles every ~10 years
  - number of journals doubles every ~15 years

The Information Explosion

1. finding relevant literature
2. accessing the literature
3. maintaining awareness of the literature
4. **reading and processing the information**

Is this a new problem?

"The individual scientist is being **overloaded with scientific information** […] and can no longer keep up with and assimilate all the information being produced"
Garvey & Griffith (1971)

How to Summarize the Results?

- traditionally:
  - narrative literature reviews
  - vote counting method
  - combining tests of significance
- now:
  - systematic reviews / meta-analysis

13

Narrative Literature Review

a description/summary of the current state of knowledge on a particular topic supported by empirical findings as well as the underlying theories and models

14

Potential Problems

- unsystematic
- subjective
- intractable
- in essence scientifically unsound

15

Non-Replicable Process

how replicable is the process of a person reading dozens or even hundreds of papers, thinking about them, and then writing down his or her conclusions?

16

Vote Counting

- examine all relevant studies conducted
- categorize based on statistical significance
  - statistically significant (with $Trt > Ctrl$ or $r > 0$)
  - not statistically significant
  - statistically significant (with $Trt < Ctrl$ or $r < 0$)
- declare most frequent category the 'winner'
- inconsistent when power of studies is low (Hedges & Olkin, 1985): as $k \to \infty$, method fails to find a true effect or association

17

Combining Tests of Significance

- long history of methods for combining the results from independent significance tests (Tippett, 1931; Fisher, 1932; Pearson, 1933; Stouffer et al., 1949; Wilkinson, 1951; Mosteller & Bush, 1954; Good, 1955; Lipták, 1958; Lancaster, 1961; …)

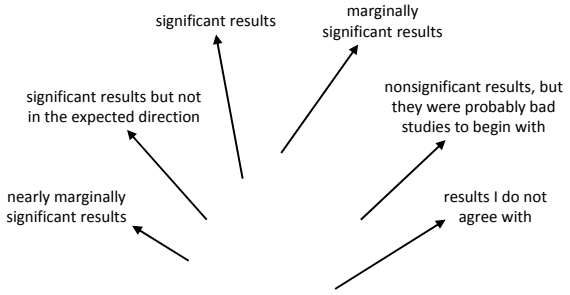
18

21·1. The Combination of Probabilities from Tests of Significance

When a number of quite independent tests of significance have been made, it sometimes happens that although few or none can be claimed individually as significant, yet the aggregate gives an impression that the probabilities are on the whole lower than would often have been obtained by chance. It is sometimes desired, taking account only of these probabilities, and not of the detailed composition of the data from which they are derived, which may be of very different kinds, to obtain a single test of the significance of the aggregate, based on the product of the probabilities individually observed (Fisher, 1932).

19

Statistical Significance



(terms used to denote results with p > .05: http://goo.gl/BRIHaa)

20

Systematic Reviews

- research synthesis as a scientific process
- based on replicable and systematic methods that are meant to "limit bias in the assembly, critical appraisal, and synthesis of all relevant studies on a specific topic" (Last, 2001)
- **methods should be made explicit**
- synthesis part can make use of qualitative or quantitative methods
- for some history, see Chalmers et al. (2002)

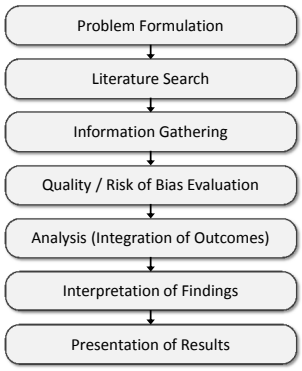
21

Meta-Analysis

- a set of statistical methods and techniques for aggregating, summarizing, and drawing inferences from collections of related studies
- **key idea:** quantify the size, direction, and/or strength of the effect or association in each study and use this as primary data in further analyses

22

Stages of a Research Synthesis



(Cooper, 2016)

23

Is Meta-Analysis More 'Objective'?

- yes and no
+ we have systematic methods
+ methods are reported
– many (subjective) decisions need to be made (researcher degrees of freedom, the 'garden of forking paths'; Gelman & Loken, 2014)
- maybe the wrong question: are MAs more transparent? YES!

24

Is Meta-Analysis More 'Objective'?

- there are MAs with conflicting conclusions (Goodyear-Smith et al., 2012; de Vrieze, 2018)
- differences often depend on inclusion criteria and handling of potential 'publication bias'

25

and / Critique

- two responses:
  - want to know the characteristics of fruits in general (fruit salad!)
  - want to examine systematic differences between various fruits
- two options:
  - a priori exclusion
  - a posteriori examination

26

Smith & Glass (1977)

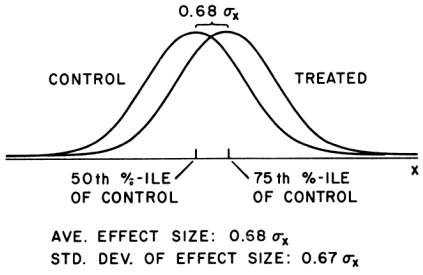


Figure 1. *Normal curves illustrating the aggregate effect of psychotherapy in relation to untreated control groups. (Data based on 833 effect size measures from 375 studies, representing about 40,000 treated and untreated subjects.)*

27

Smith & Glass (1977)

Type of therapy	Average effect size	No. of effect sizes	Standard error of mean effect size	*Mdn* treated person's percentile status in control group
Psychodynamic	.59	96	.05	72
Adlerian	.71	16	.19	76
Eclectic	.48	70	.07	68
Transactional analysis	.58	25	.19	72
Rational-emotive	.77	35	.13	78
Gestalt	.26	8	.09	60
Client-centered	.63	94	.08	74
Systematic desensitization	.91	223	.05	82
Implosion	.64	45	.09	74
Behavior modification	.76	132	.06	78

28

Some were still not convinced ...

- "mega-silliness" (Eysenck, 1978)
- "meta-analysis/shmeta-analysis" (Shapiro, 1994)
- "statistical alchemy" (Feinstein, 1995)

29

Garbage In Garbage Out Critique

- too broad: low internal validity
- too restrictive: low external validity
- two options (again):
  - a priori exclusion
  - a posteriori examination

33

The "File Drawer" Problem

- failure to obtain all relevant studies (or a representative sample thereof) on the topic of interest
- can result from only focusing on the published literature (statistically significant findings are overrepresented, results will be biased)

34

Potential Sources of Bias

- statistically significant findings are:
  - more likely to be published
  - more likely to be published quicker
  - more likely to be cited in English journals
  - more likely to be published more than once
  - more likely to be cited by others
  - …

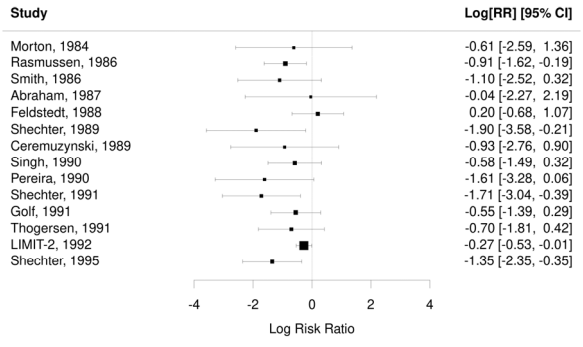
35

Example: Magnesium Treatment

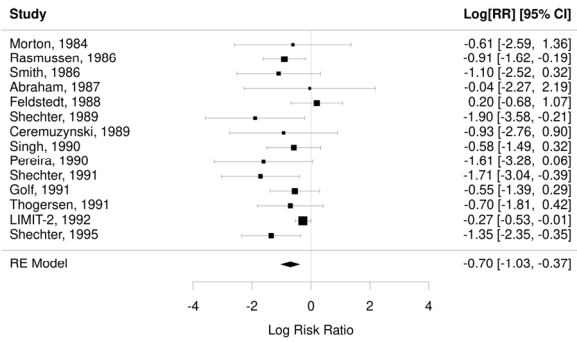
- meta-analysis on the effectiveness of intravenous magnesium treatment in acute myocardial infarction for reducing the risk of mortality and arrhythmias

37

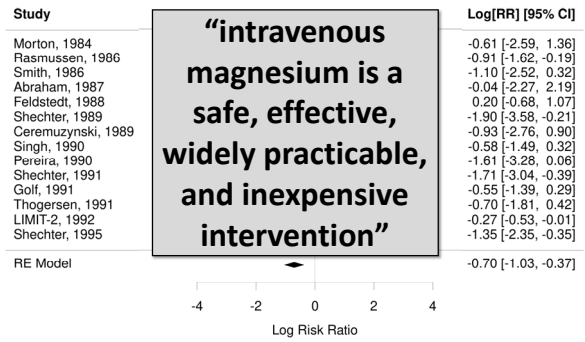
Example: Magnesium Treatment



Example: Magnesium Treatment

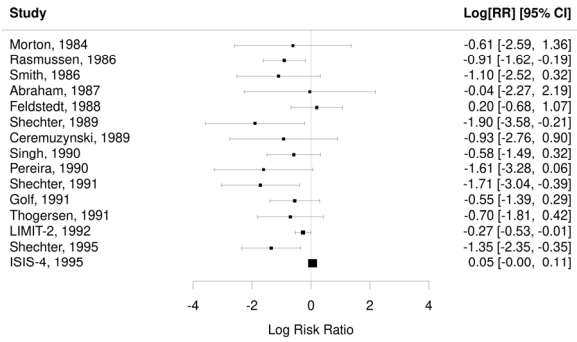


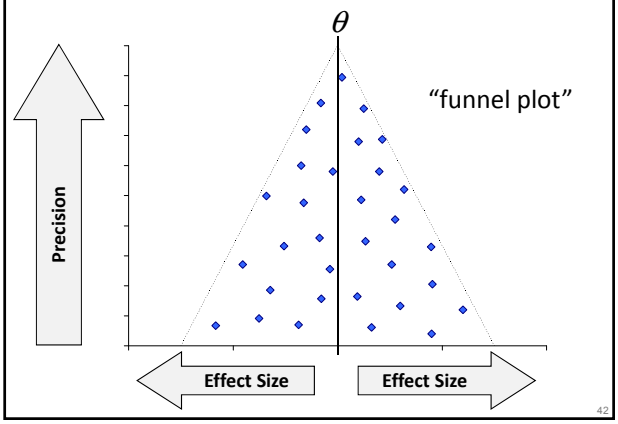
Example: Magnesium Treatment

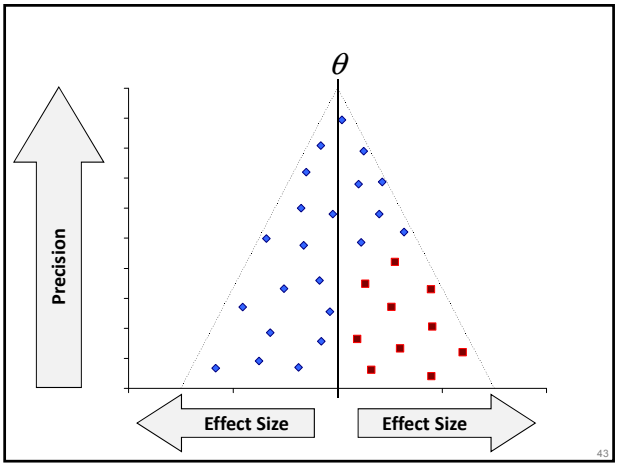


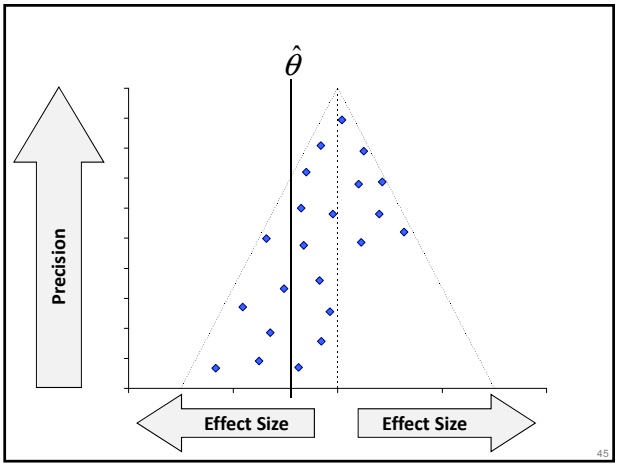
"intravenous magnesium is a safe, effective, widely practicable, and inexpensive intervention"

Example: Magnesium Treatment

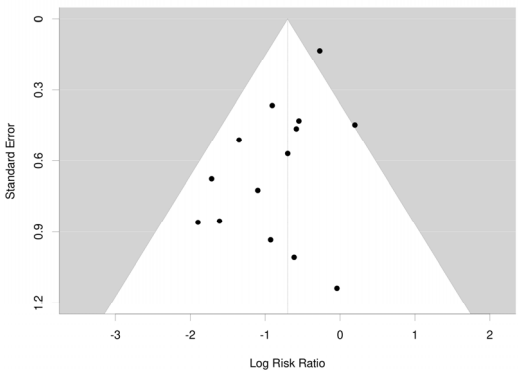








Example: Magnesium Treatment



46

Robustness to Publication Bias

- if a number of (unpublished) studies with null effects were found, they could reverse the conclusions of a meta-analysis
- how many such studies would it take?
- if this number is large, results are robust
- idea due to Rosenthal (1979), later extended by Orwin (1983) and Rosenberg (2005)
- sometimes called a 'failsafe N' calculation

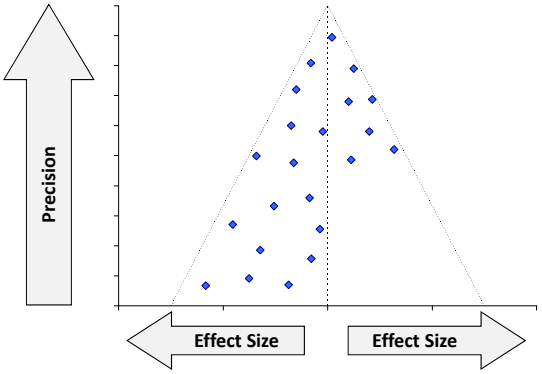
47

Robustness to Publication Bias

- Rosenthal method: 138 studies
- Orwin method: 14 studies
- Rosenberg method: 69 studies
- discrepancies due differences in underlying methods (and their purpose)
- are these numbers 'large'?
- method not used much anymore in practice

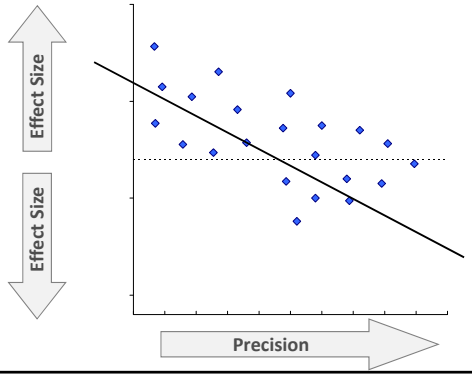
48

Regression Test



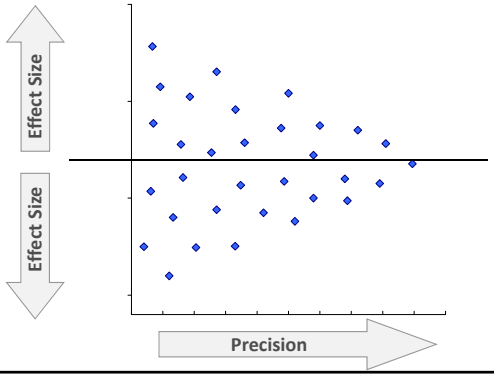
49

Regression Test



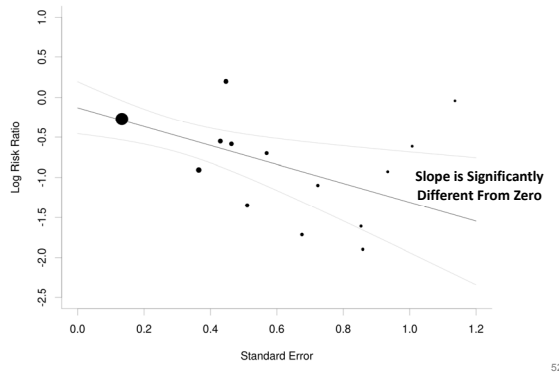
50

Regression Test



51

Example: Magnesium Treatment



52

Regression Test

- there are various versions of the regression test (all based on the same principle)
- sometimes called "Egger's test" (based on Egger et al., 1997)
- it is **test for funnel plot asymmetry**, not publication bias per se; there are many possible reasons for funnel plot asymmetry (Sterne et al., 2011; Coburn & Vevea, 2015)

53

Test of Excess Significance

- test null hypothesis of no effect in each study
- O: observed number of significant tests
- compute the power of each test, $1 - \beta_i$, given some (estimated) value of the true effect
- $E = \sum(1 - \beta_i)$: expected number of significant tests
- test if O is significantly larger than E
- Ioannidis & Trikalinos (2007)

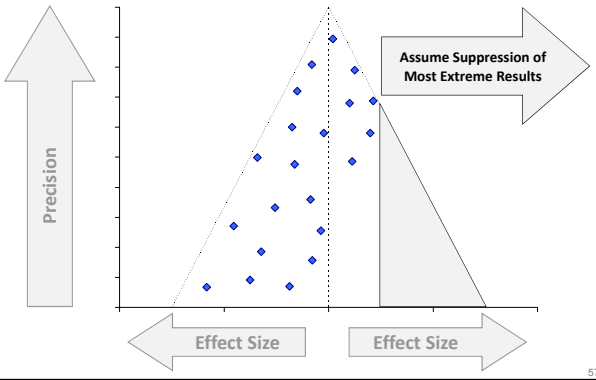
54

Example: Magnesium Treatment

- 5 significant findings in 14 studies
- power ranges from .07 to .95 (median = .12)
- expected number of significant findings: 2.71
- test of excess significance: p = 0.081

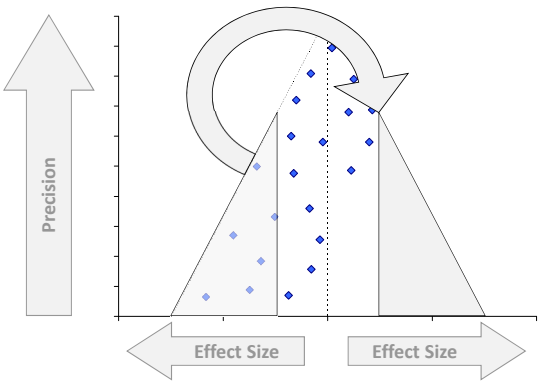
55

Trim and Fill Method



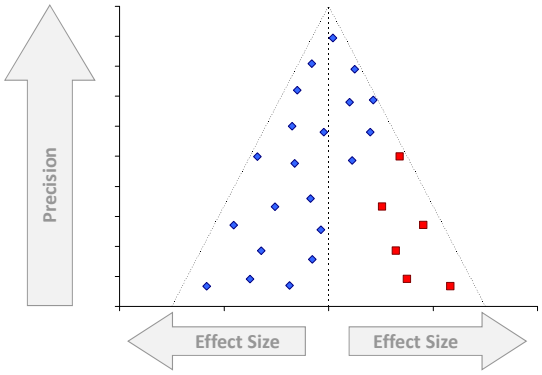
57

Trim and Fill Method



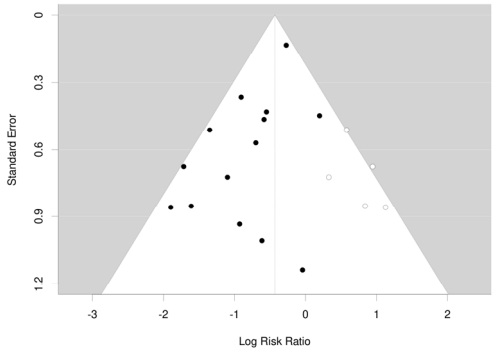
59

Trim and Fill Method



60

Example: Magnesium Treatment



61

Example: Magnesium Treatment

- meta-analysis based on the 14 studies:

  $\hat{\mu} = -0.70$ (95% CI: -1.03 to -0.37)

- trim and fill method (14 + 5 studies):

  $\hat{\mu} = -0.43$ (95% CI: -0.78 to -0.08)

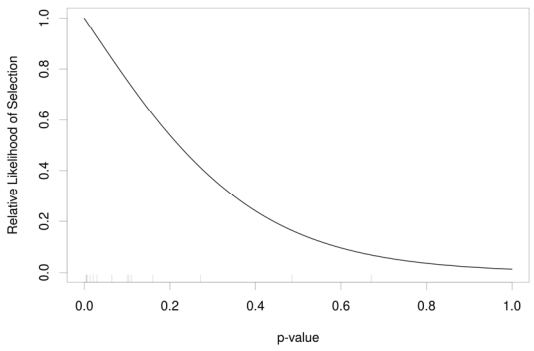
62

Selection Models

- assume there might be an inverse relationship between the p-value of each study and the probability of its inclusion in a meta-analysis
- with enough studies, can estimate this relationship and and remove the bias from the meta-analytic findings
- difficult to use in practice (models are complicated and k must be quite large)

63

Example: Magnesium Treatment



64

Example: Magnesium Treatment

- meta-analysis based on the 14 studies:

  $\hat{\mu} = -0.70$ (95% CI: -1.03 to -0.37)

- selection model:

  $\hat{\mu} = -0.32$ (95% CI: -0.59 to -0.05)

- test of selection model: p = .02

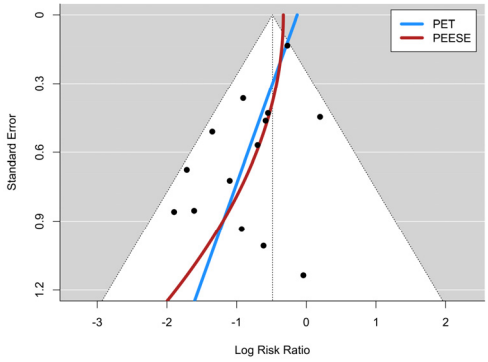
65

PET and PEESE

- PET (precision-effect test) and PEESE (precision-effect estimate with SE) are methods for estimating/testing the 'true' effect in the presence of publication bias (Stanley & Doucouliagos, 2014)
- in essence: the intercept of a 'regression test'

66

Example: Magnesium Treatment



67

Example: Magnesium Treatment

- meta-analysis based on the 14 studies:

$$\hat{\mu} = -0.70 \ (95\% \text{ CI: } -1.03 \text{ to } -0.37)$$

- PET:

$$\hat{\mu} = -0.13 \ (95\% \text{ CI: } -0.46 \text{ to } 0.20)$$

- PEESE:

$$\hat{\mu} = -0.39 \ (95\% \text{ CI: } -0.73 \text{ to } -0.06)$$

68

Publication Bias

- affects all review methods (not a problem specific to meta-analysis!)
- in fact, due to meta-analysis:
  - increased awareness of publication bias
  - development of systematic methods to detect and address publication bias
  - continued emphasis on the importance of trial registries, protocols, and registered reports (to eliminate publication bias)

69

Gold Standard

- meta-analysis of registered reports
- same as a 'prospective meta-analysis' (Simes, 1995; Berlin & Colditz, 1999)
- if not (yet) possible, acknowledge/examine the multitude of possible results (multiverse analysis; Voracek et al., 2019)

70

Some Open Science Practices for MA

- MA needs to embrace open science practices (Haddaway, 2018; Moreau & Gamble, 2022)
- write a protocol / preregistration
- provide materials (e.g., for screening, coding, data extraction)
- provide data and analysis scripts
- use free/libre open source software
- follow reporting standards (e.g., PRISMA)
- document/justify deviations from the protocol
- publish open access

71

Reporting Standards

- **QUOROM:** Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D. & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *Lancet, 354*, 1896-1900. https://doi.org/10.1159/000055014
- **MOOSE:** Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., Moher, D., Becker, B. J., Sipe, T. A. & Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *Journal of the American Medical Association, 283*(15), 2008-2012. https://doi.org/10.1001/jama.283.15.2008
- **PRISMA:** Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. & Group, T. P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Journal of Clinical Epidemiology, 62*(10), 1006-1012. https://doi.org/10.1371/journal.pmed.1000097
- **MARS:** APA Meta-Analysis Reporting Standards. American Psychological Association. URL: https://apastyle.apa.org/jars/quant-table-9.pdf

- Berlin, J. A. & Colditz, G. A. (1999). The role of meta-analysis in the regulatory process for foods, drugs, and devices. *Journal of the American Medical Association, 281*(9), 830-834. https://doi.org/10.1001/jama.281.9.830
- Chalmers, I., Hedges, L. V. & Cooper, H. (2002). A brief history of research synthesis. *Evaluation and the Health Professions, 25*(1), 12-37. https://doi.org/10.1177/0163278702025001003
- Coburn, K. M. & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods, 20*(3), 310-330. https://doi.org/10.1037/met0000046
- Cooper, H. M. (2016). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). Sage.
- Duval, S. J. & Tweedie, R. L. (2000). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98. https://doi.org/10.1080/01621459.2000.10473905
- Egger, M., Davey Smith, G., Schneider, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629-634. https://doi.org/10.1136/bmj.315.7109.629
- Fisher, R. A. (1932). *Statistical methods for research workers* (4th ed.). Oliver and Boyd.
- Fitch, V. L. (1989). An APS president reflects on his two-year term. *Physics Today, 42*(12), 49-53. https://doi.org/10.1063/1.881187
- Garvey, W. D. & Griffith, B. C. (1971). Scientific communication: Its role in the conduct of research and creation of knowledge. *American Psychologist, 26*(4), 349-362. https://doi.org/10.1037/h0032059
- Gelman, A. & Loken, E. (2014). The statistical crisis in science. *American Scientist, 102*(6), 460. https://doi.org/10.1511/2014.111.460
- Haddaway, N. R. (2018). Open synthesis: On the need for evidence synthesis to embrace open science. *Environmental Evidence, 7*(1), 26. https://doi.org/10.1186/s13750-018-0140-4
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Academic Press.
- Ioannidis, J. P. A. & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials, 4*(3), 245-253. https://doi.org/10.1177/1740774507079441
- Last, J. M. (2001). *A dictionary of epidemiology* (4th ed.). Oxford University Press.
- Moreau, D. & Gamble, B. (2022). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods, 27*(3), 426-432. https://doi.org/10.1037/met0000351
- Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics, 8*(2), 157-159. https://doi.org/10.3102/10769986008002157
- Preston, C., Ashby, D. & Smyth, R. (2004). Adjusting for publication bias: Modelling the selection process. *Journal of Evaluation in Clinical Practice, 10*(2), 313-322. https://doi.org/10.1111/j.1365-2753.2003.00457.x
- Rosenberg, M. S. (2005). The file-drawer problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution, 59*(2), 464-468. https://doi.org/10.1111/j.0014-3820.2005.tb01004.x
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641. https://doi.org/10.1037/0033-2909.86.3.638
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments.* Wiley.
- Simes, R. J. (1995). Prospective meta-analysis of cholesterol-lowering studies: The prospective pravastatin pooling (PPP) project and the cholesterol treatment trialists (CTT) collaboration. *American Journal of Cardiology, 76*(9), 122C-126C. https://doi.org/10.1016/s0002-9149(99)80482-2
- Smith, M. L. & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*(9), 752-760. https://doi.org/10.1037/0003-066x.32.9.752
- Stanley, T. D. & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60-78. https://doi.org/10.1002/jrsm.1095
- Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D. & Higgins, J. P. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 343*, d4002. https://doi.org/10.1136/bmj.d4002
- Voracek, M., Kossmeier, M. & Tran, U. S. (2019). Which data to meta-analyze, and how?. *Zeitschrift für Psychologie / Journal of Psychology, 227*(1), 64-82. https://doi.org/10.1027/2151-2604/a000357
- de Vrieze, J. (2018). The metawars. *Science, 361*(6408), 1184-1188. https://doi.org/10.1126/science.361.6408.1184