Evidence synthesis methods for investigating variation in treatment effects: Challenges and innovations

James E. Pustejovsky
University of Wisconsin - Madison

November 11, 2026

Evidence Synthesis

Evidence synthesis: The systematic integration of empirical results across multiple sources of evidence, for purposes of drawing generalizations (Cooper & Hedges, 2009).

Meta-analysis: The set of statistical methods used to conduct research synthesis, especially methods for combining results from a collection of multiple studies. The ***application*** of such methods to empirical data.

(Potential) contributions of evidence synthesis

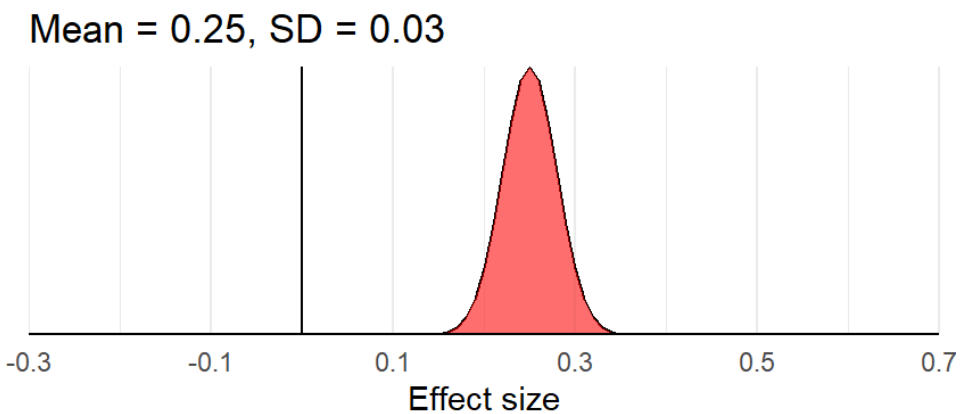
1. **Summarizing**, integrating, and clarifying patterns of findings on a topic.
2. **Understanding variation** in findings.
3. **Assessing limitations** of available evidence.
4. Informing **evidence-based practice** and **policy** guidance.

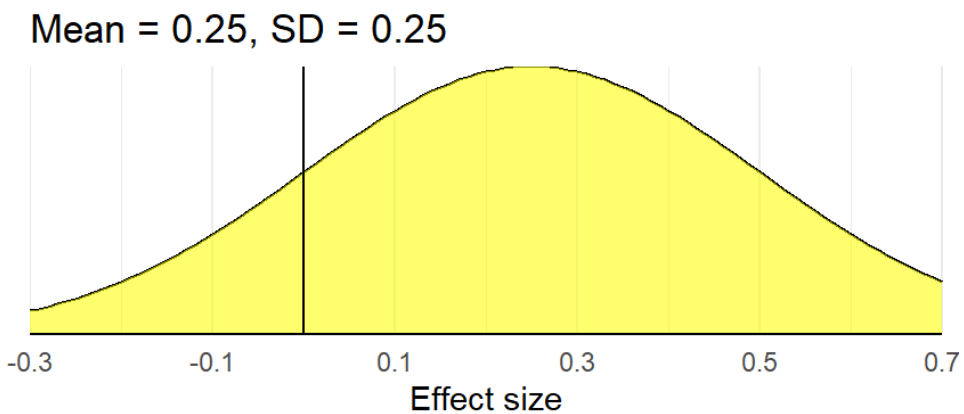
Challenges and innovations

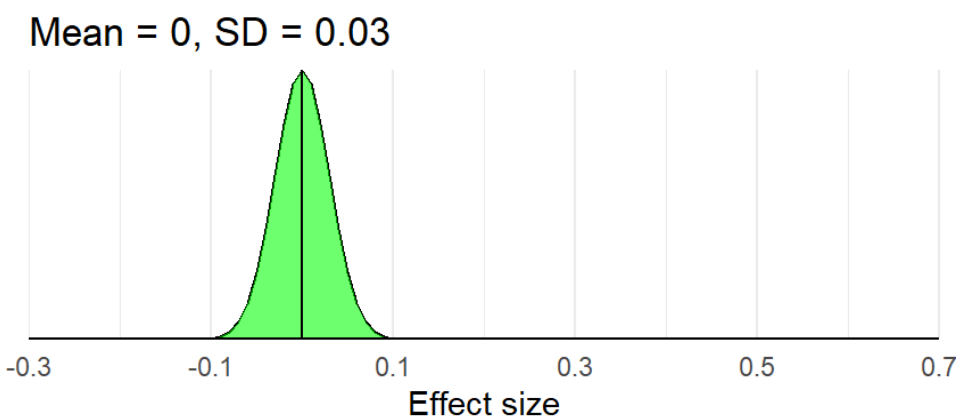
1. **Interpret the distribution**
2. Meta-analytic models are not magic
3. Moderator analysis is correlational
4. Aggregated data has limits

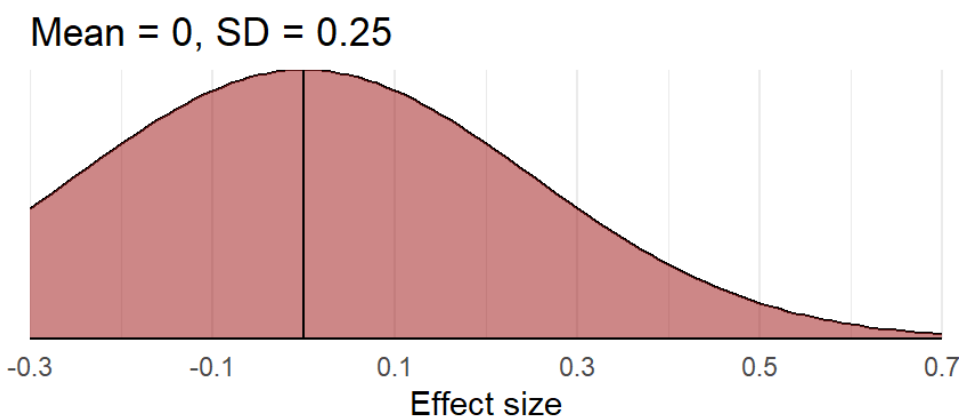
A distribution of effect size parameters

- Random effects meta-analysis assumes a **distribution** of latent effect size parameters.
  - Important to interpret the whole distribution, *not just the average effect size.*









Dependent effect sizes

Treatment	O P Q
Control	O P Q

d_{O1}
d_{P1}
d_{Q1}

Treatment	O_1	O_2	O_3
Control	O_1	O_2	O_3

d_{12}
d_{22}
d_{32}

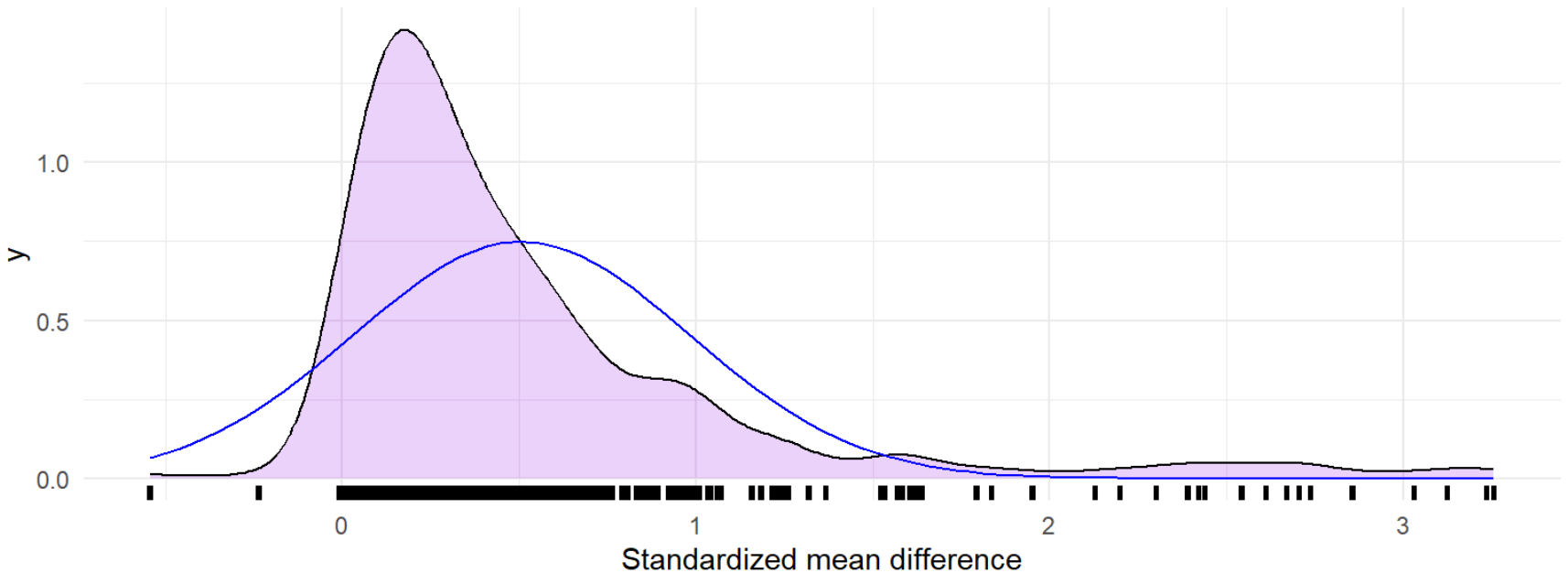
Treatment T	O
Treatment U	O
Control	O

d_{T3}
d_{U3}

- Dependent effect sizes are ubiquitous in psychology and social science meta-analyses.
- Effect sizes have a hierarchical structure, with multiple estimates nested within a study.
- Meta-analytic models can describe the **amount** and **structure** of heterogeneity.

Mindful movement synthesis

Deng et al. (2025) examined the effects of *mindful movement* on *cognitive functioning* of school-age children.



- Mean of effect size distribution: $\hat{\mu} = 0.507$, 95% CI [0.408, 0.605]
- SDs of effect size distribution: $\hat{\tau}_{total} = 0.475$, $\hat{\tau}_{between} = 0.375$, $\hat{\tau}_{within} = 0.292$.
- 80% prediction interval for effect size in a new study: $[-0.105, 1.119]$

Challenges and innovations

1. Interpret the distribution

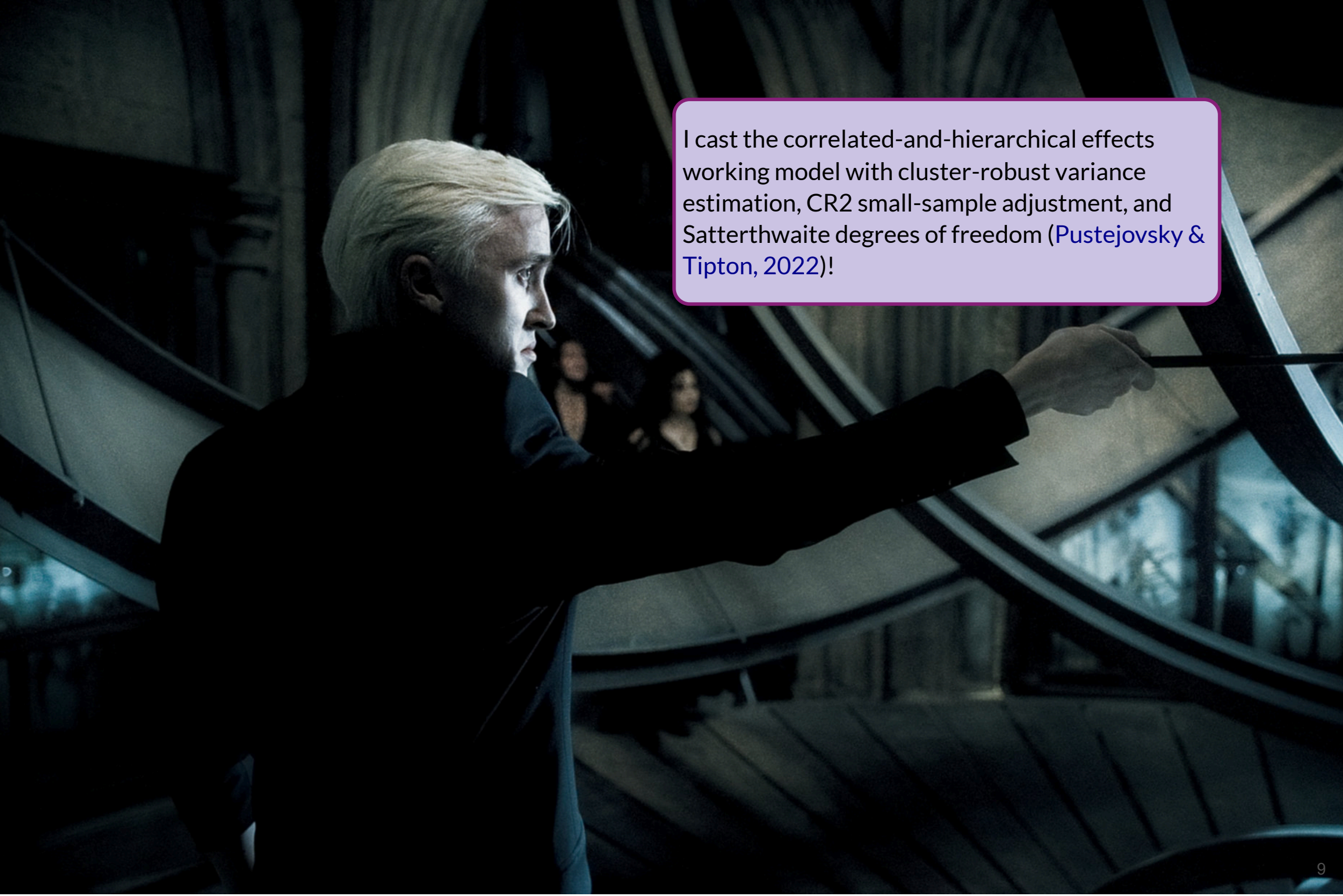
2. Meta-analytic models are not magic

3. Moderator analysis is correlational

4. Aggregated data has limits

More complex models for more complex data

- Hierarchical meta-analysis (Konstantopoulos, 2011; Van den Noortgate et al., 2013)
- Correlated-and-hierarchical effects models (Pustejovsky & Tipton, 2022)
- Multivariate, multilevel meta-analysis (McShane & Böckenholt, 2018; Sera et al., 2019)
- Network meta-analysis (Higgins & Whitehead, 1996; Lu & Ades, 2006)
- Robust Bayesian model-averaged meta-analysis (Bartoš et al., 2023, 2026; Maier et al., 2023)


I cast the correlated-and-hierarchical effects working model with cluster-robust variance estimation, CR2 small-sample adjustment, and Satterthwaite degrees of freedom (Pustejovsky & Tipton, 2022)!


I cast Robust Bayesian Model-Averaged Meta-Analysis with Selection Models (Bartoš et al., 2023)!

Preliminary data analysis methods

- Initial data analysis as a formal process (Baillie et al., 2022; Heinze et al., 2024; Huebner et al., 2018; Lusa et al., 2024).

PReliminary Investigation of MEta-analytic Databases (PRIMED, Pustejovsky et al., 2026)

Describe the amount of data and its dependence structure

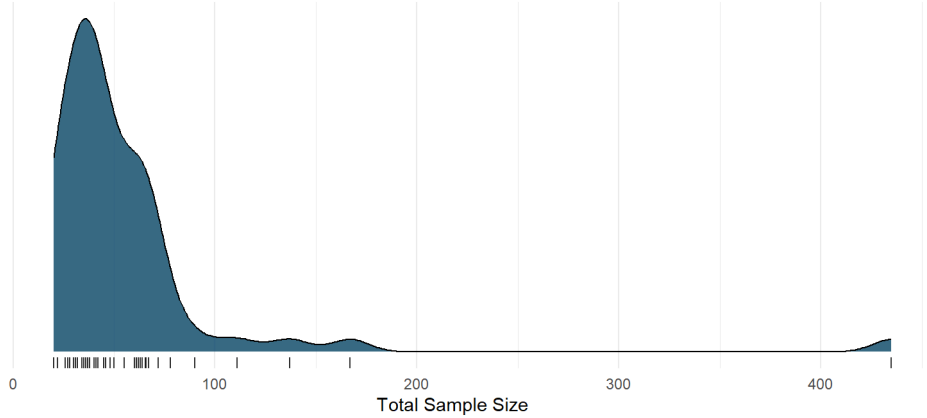
Explore study characteristics and potential moderators

Inspect standard errors and other auxiliary data

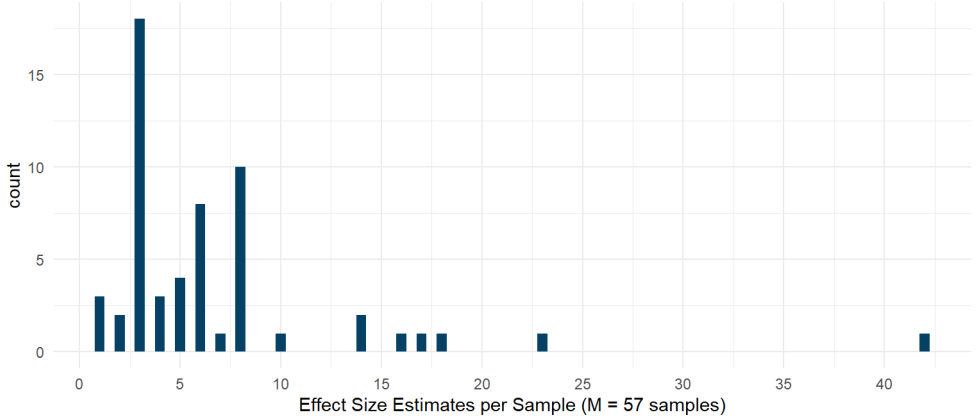
Visualize the distribution of effect size estimates

Describe the amount of data and its dependence structure

Sample size distribution



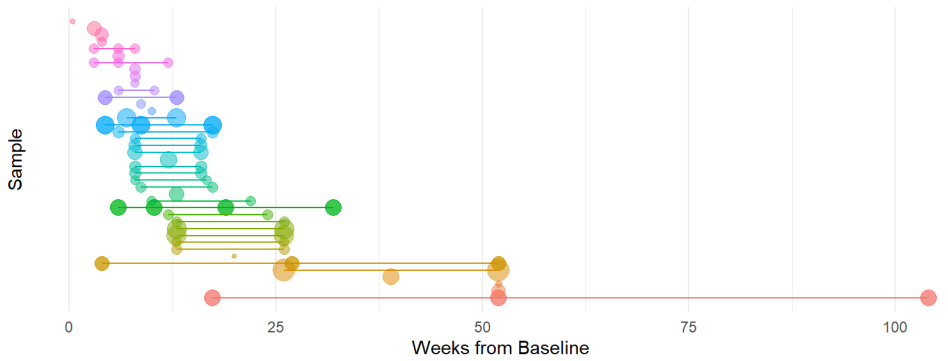
Effect size estimates per sample



Source: Winters et al. (2022). Investigating narrative performance in children with developmental language disorder: A systematic review and meta-analysis.

Explore study characteristics and potential moderators

Follow-up times vary within and between studies



Source: McLouth et al. (2021). A systematic review and meta-analysis of effects of psychosocial interventions on spiritual well-being in adults with cancer.

Within-study co-occurrence of assessment types

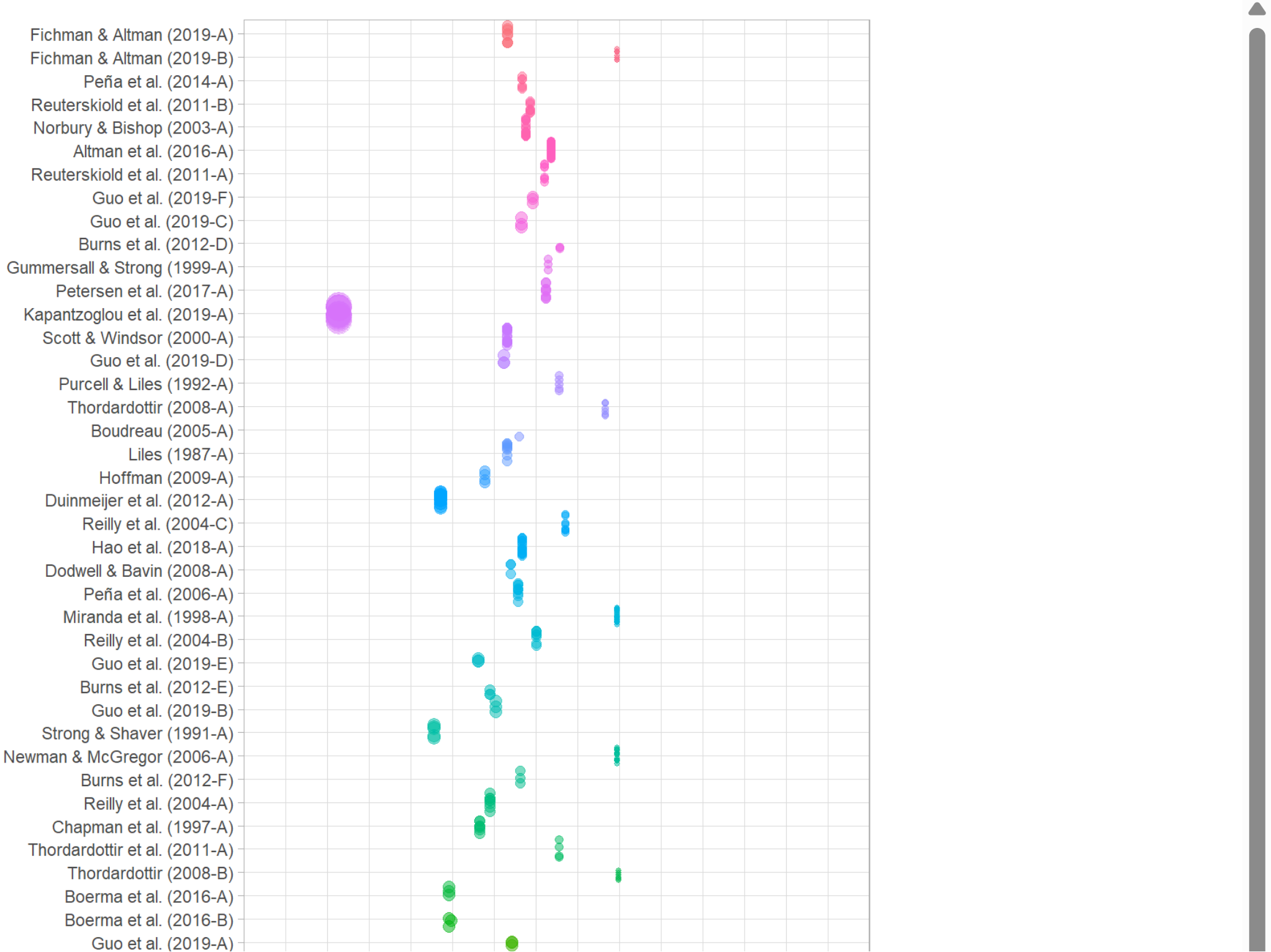
Assessment type	(A)	(B)	(C)	(D)
(A) ISL	3 (4, 20)	3 (4, 20)	2 (2, 18)	
(B) Macro	3 (4, 13)	26 (39, 111)	22 (33, 87)	4 (11, 1
(C) Micro	2 (2, 31)	22 (33, 153)	31 (49, 235)	4 (11, 1
(D) Mixed		4 (11, 14)	4 (11, 14)	6 (13, 1

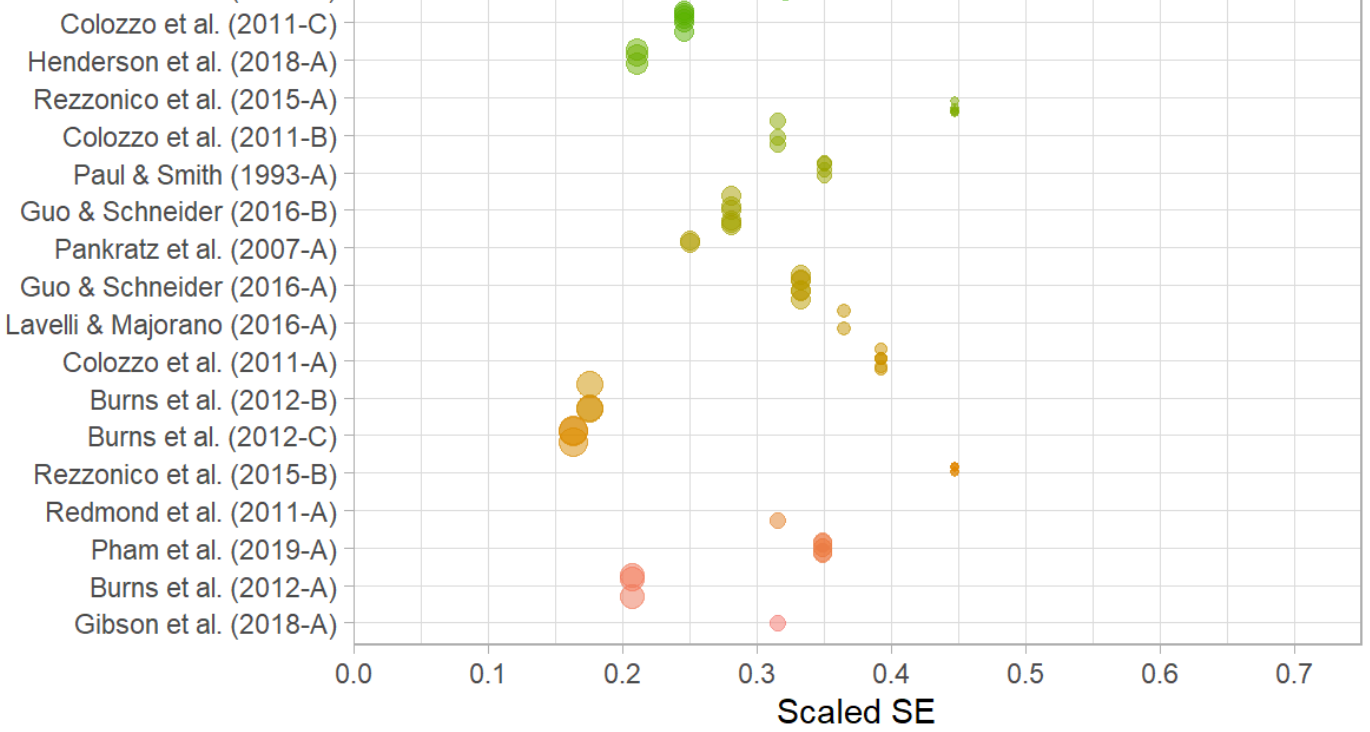
Values outside parentheses indicate the number of studies;

Values in parentheses indicate the number of samples and number of effect size estimates.

Source: Winters et al. (2022). Investigating narrative performance in children with developmental language disorder: A systematic review and meta-analys

Inspect standard errors and other auxiliary data


Fichman & Altman (2019-A)
Fichman & Altman (2019-B)
Peña et al. (2014-A)
Reuterskiold et al. (2011-B)
Norbury & Bishop (2003-A)
Altman et al. (2016-A)
Reuterskiold et al. (2011-A)
Guo et al. (2019-F)
Guo et al. (2019-C)
Burns et al. (2012-D)
Gummersall & Strong (1999-A)
Petersen et al. (2017-A)
Kapantzoglou et al. (2019-A)
Scott & Windsor (2000-A)
Guo et al. (2019-D)
Purcell & Liles (1992-A)
Thordardottir (2008-A)
Boudreau (2005-A)
Liles (1987-A)
Hoffman (2009-A)
Duinmeijer et al. (2012-A)
Reilly et al. (2004-C)
Hao et al. (2018-A)
Dodwell & Bavin (2008-A)
Peña et al. (2006-A)
Miranda et al. (1998-A)
Reilly et al. (2004-B)
Guo et al. (2019-E)
Burns et al. (2012-E)
Guo et al. (2019-B)
Strong & Shaver (1991-A)
Newman & McGregor (2006-A)
Burns et al. (2012-F)
Reilly et al. (2004-A)
Chapman et al. (1997-A)
Thordardottir et al. (2011-A)
Thordardottir (2008-B)
Boerma et al. (2016-A)
Boerma et al. (2016-B)
Guo et al. (2019-A)

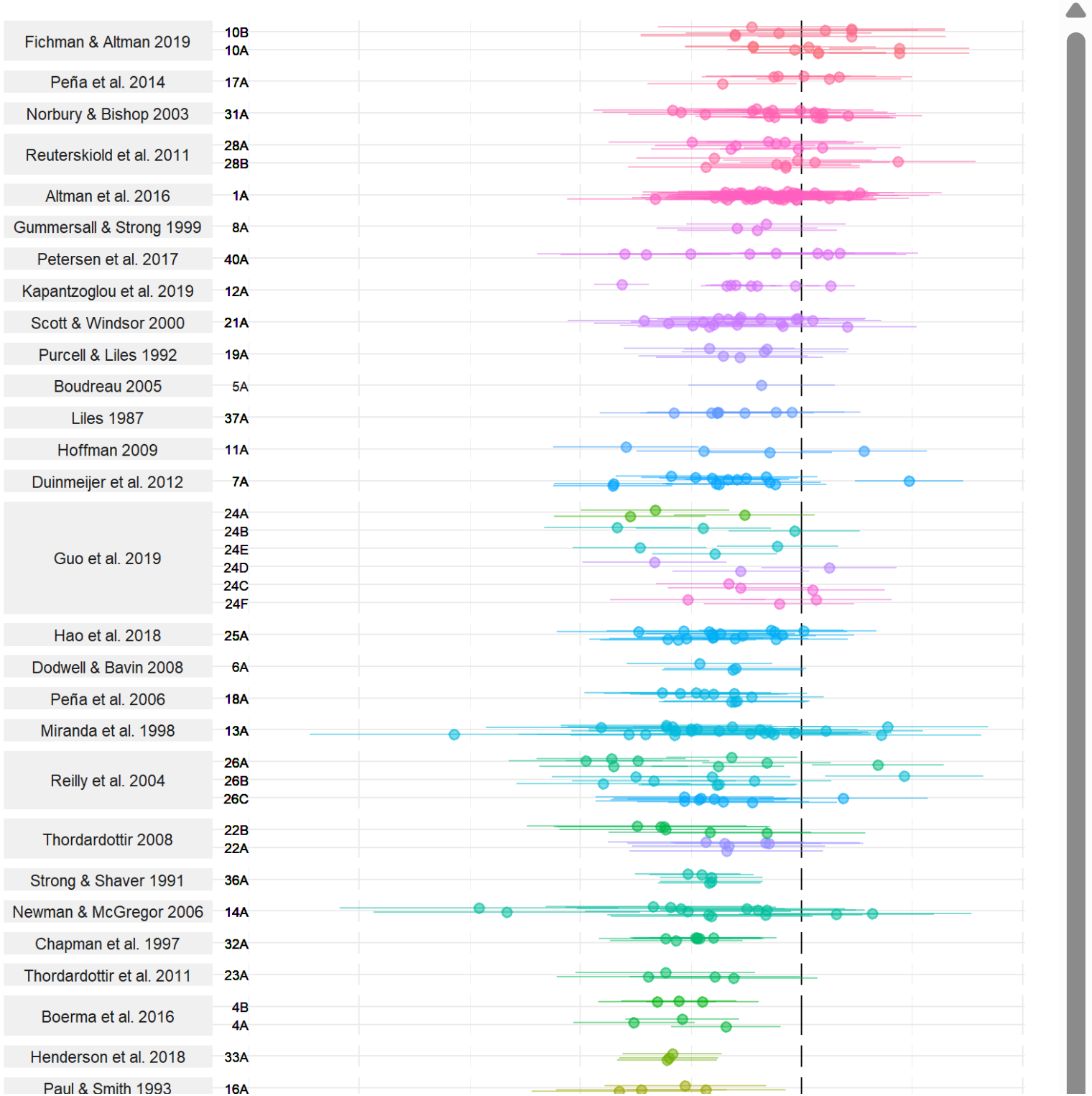


Source: Winters et al. (2022). Investigating narrative performance in children with developmental language disorder: A systematic review and meta-analysis.

Visualize the distribution of effect size estimates

Hierarchical forest plot

Source: Winters et al. (2022). Investigating narrative performance in children with developmental language disorder: A systematic review and meta-analysis.



Challenges and innovations

1. Interpret the distribution
2. Meta-analytic models are not magic
3. **Moderator analysis is correlational**
4. Aggregated data has limits

What works for whom, how, and under what conditions?

- **Moderator analysis** is used to investigate variation in effect sizes.
  - How are study characteristics related to magnitude of intervention effects?
  - Who benefits? How are benefits and harms distributed across population of interest?

- Moderator analyses often examine variation in effect size based on characteristics of primary study participants and contexts:
  - Participants' family income level
  - Presence or absence of intervention features
  - Study setting
  - Type/category of outcome measure

Moderator analysis through meta-regression:

- Meta-regression with a continuous moderator X_1:

$$T_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

  - β_1 measures differences in average effects for studies differing by 1 unit on X_1

- Or for a categorical predictor with levels A, B, C, D:

$$T_i = \beta_0 + \beta_1 (X_{1i} = B) + \beta_2 (X_{1i} = C) + \beta_3 (X_{1i} = D) + \epsilon_i$$

Are there differences by domain of cognitive functioning?

Average intervention duration (hours) by cognitive domain

domain_	M	SD	min	median	max
A	11.94	14.07	1	8.00	78
EF	15.73	16.38	1	12.00	88
WM	9.79	5.97	1	10.50	24
PS	14.25	10.75	1	16.00	24

Distinguishing direct versus contextual evidence

Direct evidence

- Reported effect size estimates for each of multiple subgroups, intervention types, outcome types, etc..
- Study-level operational features are held constant.
- Can provide estimates of ***individual-level variation*** in impacts.

Contextual evidence

- **Sample-level average** effect size estimates and **average sample characteristics**.
- Open to **aggregation bias** (a.k.a. the ecological fallacy).

Direct and contextual evidence are *conceptually distinct*...

...and should be analyzed as such.

- Meta-analyze the direct evidence (co-occurring effect sizes) alone, excluding the contextual evidence.

and/or

- Center the predictor by sample, include the centered predictor and the sample-level averaged predictor in a meta-regression.

Meta-analyze the direct evidence alone

- Analyze the direct evidence (co-occuring effect sizes) in a separate meta-analysis, excluding the contextual evidence.

$$\begin{pmatrix} ES_j^{WM} \\ ES_j^{A} \end{pmatrix} = \begin{pmatrix} \mu_{WM} \\ \mu_{A} \end{pmatrix} + \begin{pmatrix} v_{0j} \\ v_{1j} \end{pmatrix} + \begin{pmatrix} e_{0j} \\ e_{1j} \end{pmatrix}$$

Center by sample

- Calculate sample-level aggregate characteristic for each unique sample:

$$\left(\overline{Attention}\right)_j = \frac{1}{\sum_{i=1}^{k_j} N_{ij}} \sum_{i=1}^{k_j} N_{ij} \times \left(Attention\right)_{ij}$$

Challenges and innovations

1. Interpret the distribution
2. Meta-analytic models are not magic
3. Moderator analysis is correlational

4. Aggregated data has limits

Conventional meta-analysis uses summary results

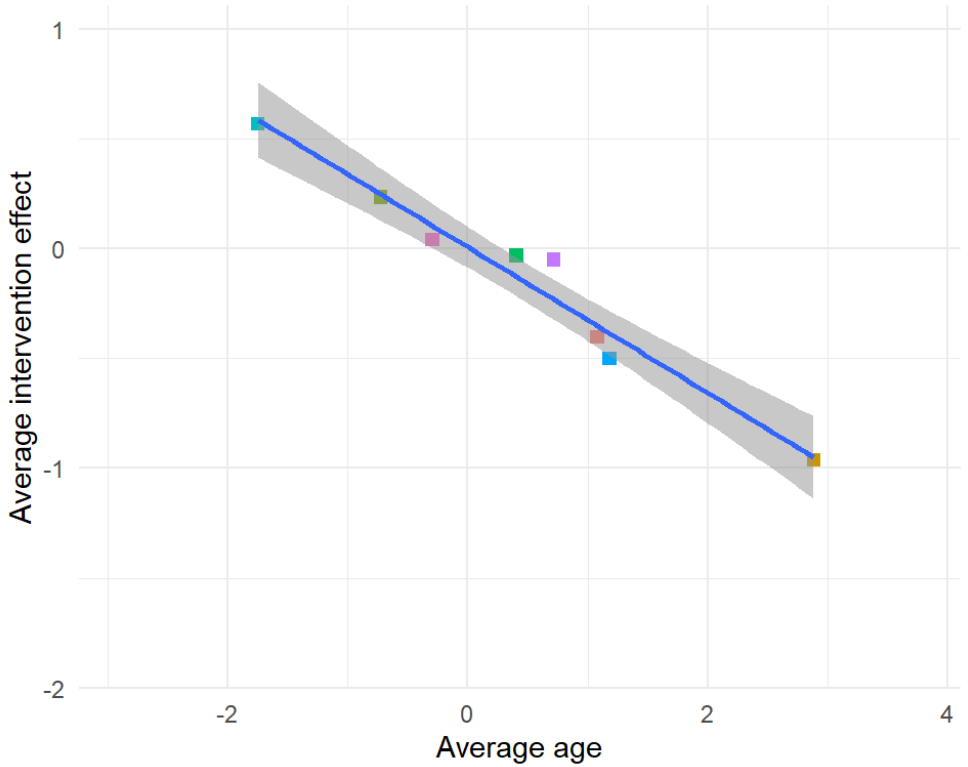
- Effect size estimates in meta-analysis are usually **summaries** or **averages** across individual participants.
  - Each observation is an average intervention effect across participants in a primary study
  - Studies differ in sample composition.

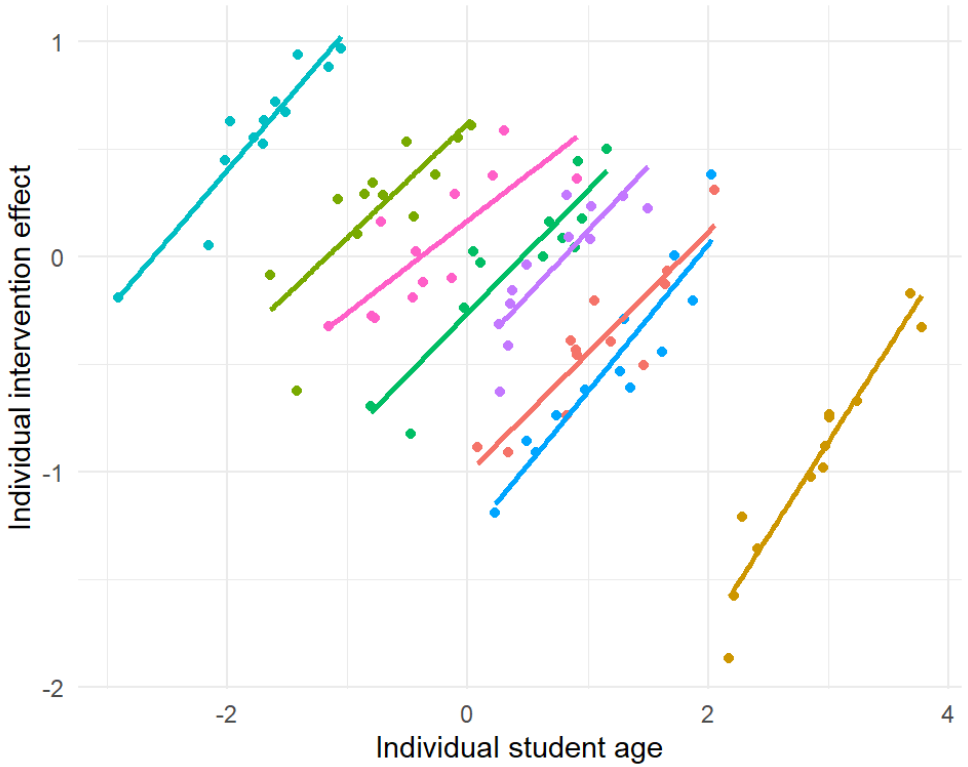
- We could use moderator analysis to understand how intervention effects vary in relation to sample characteristics:

$$T_i = \beta_0 + \beta_1(\text{Average age})_i + \epsilon_i$$

- But **an aggregate-level relation does not imply an individual-level relation**.
  - The "ecological fallacy"
  - Aggregation bias

A stylized example





Within-study relation differs from aggregate-level relation (Simpson's paradox).

Individual participant data synthesis

1. Get the raw data from the primary studies.
2. Analyze all of it in one big (multilevel) model.

Raw data allows **analysis of individual-level moderation effects** and even **testing for aggregation bias** (Lambert et al., 2002; Schmid et al., 2004).

We can **adjust for differences in sample composition** and conduct **targeted generalizations**, using tools from causal inference (Dahabreh et al., 2020).

Challenges in IPD synthesis

- Difficult and time-consuming to obtain IPD (Nevitt et al., 2017; Polanin, 2018; Polanin & Terzian, 2019).
  - But increasing use of registered reports and multilab replication studies helps!
- IPD is usually only available for a *subset* of primary studies.
  - Potential for availability bias (Ohmann et al., 2021; Riley et al., 2021).
  - But it is possible to jointly model IPD and AD (Riley et al., 2008; Zhang & Pustejovsky, 2026).
- Need to **harmonize** variables measured in many different ways.
  - But IPD allows one to evaluate different approaches and effect metrics (Cooper & Patall, 2009; Tierney et al., 2015).

Challenges and innovations

1. Interpret the distribution
2. Meta-analytic models are not magic
3. Moderator analysis is correlational
4. Aggregated data has limits

References

Baillie, M., Le Cessie, S., Schmidt, C. O., Lusa, L., Huebner, M., & for the Topic Group "Initial Data Analysis" of the STRATOS Initiative. (2022). Ten simple rules for initial data analysis. *PLOS Computational Biology*, *18*(2), e1009819. https://doi.org/10.1371/journal.pcbi.1009819

Bartoš, F., Maier, M., & Wagenmakers, E.-J. (2026). Robust Bayesian multilevel meta-analysis: Adjusting for publication bias in the presence of dependent effect sizes. *Behavior Research Methods*, *58*(6), 165. https://doi.org/10.3758/s13428-026-03023-y

Bartoš, F., Maier, M., Wagenmakers, E.-J., Doucouliagos, H., & Stanley, T. D. (2023). Robust Bayesian meta-analys Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods*, *14*(1), 99–116. https://doi.org/10.1002/jrsm.1594

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant da versus aggregated data. *Psychological Methods*, *14*(2), 165–176. https://doi.org/10.1037/a0015565

Dahabreh, I. J., Petito, L. C., Robertson, S. E., Hernán, M. A., & Steingrimsson, J. A. (2020). Toward causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a new target population. *Epidemiology (Cambridge, Mass.)*, *31*(3), 334–344. https://doi.org/10.1097/EDE.0000000000001177

Deng, Y., Zhang, Y., Zhang, J., & Huang, S. (2025). The efficacy of mindful movement for improving cognitive functioning among children and adolescents: A systematic review and meta-analysis. *Developmental Review 78*, 101229. https://doi.org/10.1016/j.dr.2025.101229

Heinze, G., Baillie, M., Lusa, L., Sauerbrei, W., Schmidt, C. O., Harrell, F. E., & Huebner, M. (2024). Regression without regrets –initial data analysis is a prerequisite for multivariable regression. *BMC Medical Research Methodology*, *24*(1, 1), 1–17. https://doi.org/10.1186/s12874-024-02294-3

Higgins, J. P. T., & Whitehead, A. (1996). Borrowing strength from external trials in a meta-analysis. *Statistics in Medicine*, *15*(24), 2733–2749. https://doi.org/10.1002/(SICI)1097-0258(19961230)15:24<2733::AID-SIM562>3.0.CO;2-0

Huebner, M., Le Cessie, S., Schmidt, C. O., & Vach, W. (2018). A contemporary conceptual framework for initial data analysis. *Observational Studies*, *4*(1), 171–192. https://doi.org/10.1353/obs.2018.0014

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis: Three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76. https://doi.org/10.1002/jrsm.35

Lambert, P. C., Sutton, A. J., Abrams, K. R., & Jones, D. R. (2002). A comparison of summary patient-level covariat in meta-regression with individual patient data meta-analysis. *Journal of Clinical Epidemiology*, *55*(1), 86–94 https://doi.org/10.1016/s0895-4356(01)00414-0

Lu, G., & Ades, A. E. (2006). Assessing evidence inconsistency in mixed treatment comparisons. *Journal of the American Statistical Association*, *101*(474), 447–459. https://doi.org/10.1198/016214505000001302

Lusa, L., Proust-Lima, C., Schmidt, C. O., Lee, K. J., Cessie, S. le, Baillie, M., Lawrence, F., Huebner, M., & Initiative, on behalf of T. of the S. (2024). Initial data analysis for longitudinal studies to build a solid foundation for reproducible analysis. *PLOS ONE*, *19*(5), e0295726. https://doi.org/10.1371/journal.pone.0295726

Maier, M., Bartoš, F., & Wagenmakers, E.-J. (2023). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *Psychological Methods*, *28*(1), 107–122. https://doi.org/10.1037/met0000405

McLouth, L. E., Ford, C. G., Pustejovsky, J. E., Park, C. L., Sherman, A. C., Trevino, K., & Salsman, J. M. (2021). A systematic review and meta-analysis of effects of psychosocial interventions on spiritual well-being in adul with cancer. *Psycho-Oncology*, *30*(2), 147–158. https://doi.org/10.1002/pon.5562

McShane, B. B., & Böckenholt, U. (2018). Multilevel multivariate meta-analysis with application to choice overload. *Psychometrika*, *83*(1), 255–271. https://doi.org/10.1007/s11336-017-9571-z

Nevitt, S. J., Marson, A. G., Davie, B., Reynolds, S., Williams, L., & Smith, C. T. (2017). Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: Systematic review. *BMJ (Clinical Research Ed.)*, *357*, j1390. https://doi.org/10.1136/bmj.j1390

Ohmann, C., Moher, D., Siebert, M., Motschall, E., & Naudet, F. (2021). Status, use and impact of sharing individu participant data from clinical trials: A scoping review. *BMJ Open*, *11*(8), e049228. https://doi.org/10.1136/bmjopen-2021-049228

Polanin, J. R. (2018). Efforts to retrieve individual participant data sets for use in a meta-analysis result in moderate data sharing but many data sets remain missing. *Journal of Clinical Epidemiology*, *98*, 157–159. https://doi.org/10.1016/j.jclinepi.2017.12.014

Polanin, J. R., & Terzian, M. (2019). A data-sharing agreement helps to increase researchers' willingness to share primary data: Results from a randomized controlled trial. *Journal of Clinical Epidemiology*, *106*, 60–69. https://doi.org/10.1016/j.jclinepi.2018.10.006

Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with robust variance estimation: Expanding the range of working models. *Prevention Science*, *23*(3), 425–438. https://doi.org/10.1007/s11121-021-01246-3

Pustejovsky, J. E., Zhang, J., & Tipton, E. (2026). A preliminary data analysis workflow for meta-analysis of dependent effect sizes. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineerir Sciences*, *384*(2321), 20240604. https://doi.org/10.1098/rsta.2024.0604

Riley, R. D., Lambert, P. C., Staessen, J. A., Wang, J., Gueyffier, F., Thijs, L., & Boutitie, F. (2008). Meta-analysis of continuous outcomes combining individual patient data and aggregate data. *Statistics in Medicine*, *27*(11), 1870–1893. https://doi.org/10.1002/sim.3165

Riley, R. D., Stewart, L. A., & Tierney, J. F. (2021). Individual participant data meta-analysis for healthcare researc In *Individual Participant Data Meta-Analysis* (pp. 1–6). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119333784.ch1

Schmid, C. H., Stark, P. C., Berlin, J. A., Landais, P., & Lau, J. (2004). Meta-regression detected associations betwe heterogeneous treatment effects and study-level, but not patient-level, factors. *Journal of Clinical Epidemiology*, *57*(7), 683–697. https://doi.org/10.1016/j.jclinepi.2003.12.001

Sera, F., Armstrong, B., Blangiardo, M., & Gasparrini, A. (2019). An extended mixed-effects framework for meta-analysis. *Statistics in Medicine*, *38*(29), 5429–5444. https://doi.org/10.1002/sim.8362

Tierney, J. F., Vale, C., Riley, R., Smith, C. T., Stewart, L., Clarke, M., & Rovers, M. (2015). Individual participant dat (IPD) meta-analyses of randomised controlled trials: Guidance on their use. *PLOS Medicine*, *12*(7), e100185 https://doi.org/10.1371/journal.pmed.1001855

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. https://doi.org/10.3758/s13428-012-0261-6

Winters, K. L., Jasso, J., Pustejovsky, J. E., & Byrd, C. T. (2022). Investigating narrative performance in children with developmental language disorder: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *65*(10), 3908–3929. https://doi.org/10.1044/2022_JSLHR-22-00017