Four Things Every Quantitative Social Scientist Should Know about Meta-Analysis

James E. Pustejovsky

Educational Psychology, PIE Colloquium

September 27, 2021

~~Things That Make Me Grumpy When I Review Research Syntheses~~

James E. Pustejovsky

Educational Psychology, PIE Colloquium

September 27, 2021

Research Synthesis and Meta-Analysis

Research 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 research 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.

Outline

- Motivating examples and data structures.
- Meta-analyses are about distributions.
- Inclusion criteria need to strike a balance.
- Moderator analysis is correlational.
- Aggregated data has limits.

Epistemic cognition interventions

Cartiff, Duke, & Greene (2021) examined effects of epistemic cognition interventions on academic performance.

- Effect sizes were standardized mean differences measuring effects of educational interventions targeting epistemic cognition or epistemic beliefs on various domains of academic performance (argumentation, conceptual, declarative, procedural, or general knowledge).
- $k = 59$ effect sizes from 28 studies.

Success for All

Cheung and colleagues (2021) reported a synthesis of Success for All, a comprehensive whole-school intervention for increasing reading abilities of elementary students.

- Effect sizes were standardized mean differences measuring intervention effects on reading performance.
- 17 randomized trials or quasi-experimental studies conducted between 1994 and 2015.

Meta-analytic data

For each of the primary studies included in a synthesis, we have:

- One or more effect size estimates,
- Standard errors for each effect size, and
- One or more predictor variables that encode study characteristics.

Sample	ES estimate	Std. Error	X1i	X2i	X3i
Angeli, 1999	0.580	0.2406	Immediate	experimental	Short term
Angeli, 1999	0.940	0.2484	Immediate	experimental	Short term
Barger et al., 2018	-0.020	0.1268	Immediate	quasi-experimental	Semester
Barzilai & Ka'adan, 2017	0.120	0.2465	Immediate	quasi-experimental with equated groups	Short term
Barzilai & Ka'adan, 2017	0.550	0.2489	Immediate	quasi-experimental with equated groups	Short term

Dependent effect size estimates

Multiple outcomes measured on a common set of participants

Treatment	O P Q
Control	O P Q

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

Outcomes measured at multiple follow-up times

Treatment	O_1	O_2	O_3
Control	O_1	O_2	O_3

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

Multiple treatment conditions compared to a common control

Treatment T	O
Treatment U	O
Control	O

d_{T3}
d_{U3}

- Dependent effect sizes are **very common** in contemporary meta-analyses.
- Effect sizes have a hierarchical structure, with multiple estimates nested within a study.
- This leads to some statistical challenges for conducting meta-analysis.

Meta-analyses are about *distributions*

From the abstract of Cartiff, Duke, & Green (2021):

> Using 28 independent samples and 59 effect sizes, we found epistemic cognition interventions had a statistically significant, medium-level effect on academic achievement (Cohen's $d = 0.509, p < .001$).

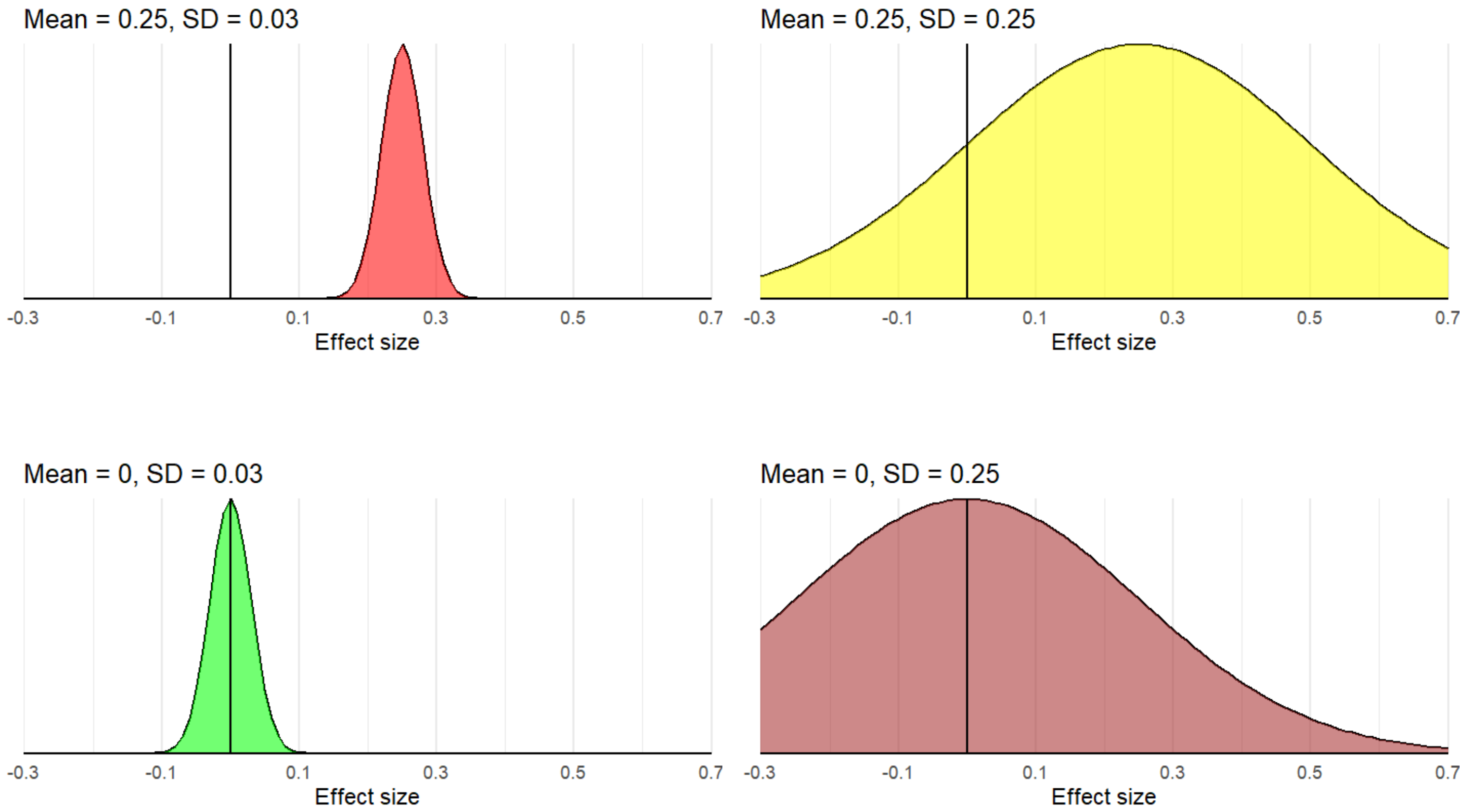
A distribution of effect sizes

- Meta-analytic models typically assume a **distribution** of true effect sizes.
    - Each study has a true effect size parameter, θ_i, which isn't observed directly.
    - We observe T_i, a *noisy estimate* of θ_i.
    - True effect sizes follow a distribution with mean μ and standard deviation τ:

$$\theta_i \sim N\left(\mu, \tau\right)$$

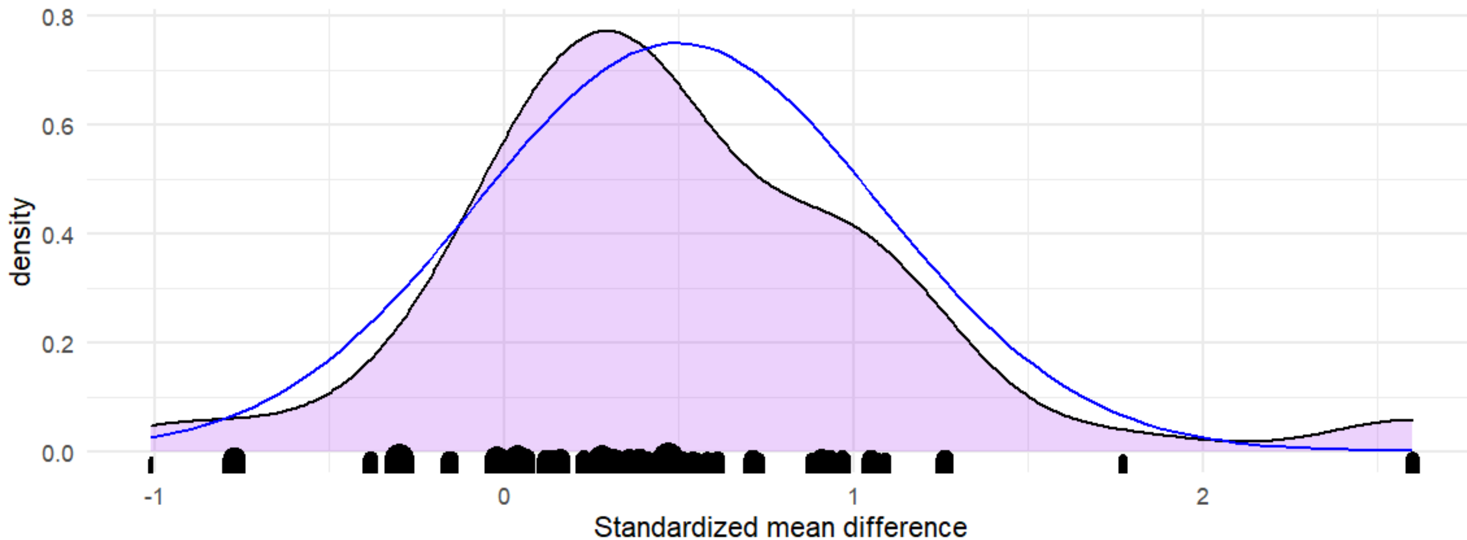
- Meta-analysis provides estimates of μ and τ.
- ***Interpret the whole distribution, not just the average effect size***.

Hypothetical effect size distributions



Epistemic cognition interventions

Distribution of effect size estimates (purple) and estimated random effects distribution (blue) from meta-analysis of epistemic cognition interventions



- Mean of effect size distribution: $\hat{\mu} = 0.497$, 95% CI [0.294, 0.700]
- SD of effect size distribution: $\hat{\tau} = 0.577$

Tools

- Report/look for estimates of τ (or τ^2).
- Prediction intervals for a new effect size.
    - 80% prediction interval for epistemic cognition interventions: [-0.250, 1.244]

Challenges

- Need further development of prediction intervals
    - for handling dependent effect sizes, hierarchical variation
    - that are more robust to non-normality

Inclusion criteria need to strike a balance

Striking a balance

- Research syntheses involve specifying a prior **inclusion** and **exclusion** criteria.
- **Generalization** is supported when a pattern of effects is found across multiple operational representations of a construct.
  - Broad inclusion criteria potentially allow for broader generalization.
- But operations often vary in relevance and quality.
  - Overly inclusive or indiscriminative criteria will inflate heterogeneity, possibly introduce bias.

> A mass of reports—good, bad, and indifferent—are fed into the computer in the hope that people will cease caring about the quality of the material on which the conclusions are based. If their abandonment of scholarship were to be taken seriously, a daunting but improbable likelihood, it would mark the beginning of a passage into the dark age of scientific psychology.

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33*(5), 517.

Clear, critical inclusion criteria

Cheung and colleagues (2021) reported a synthesis of Success for All, a comprehensive whole-school intervention for increasing reading abilities of elementary students.

- **Participants**: Students in pre-kindergarten, kindergarten, or 1st grade
- **Intervention**: Success for All program, with minimum duration of 1 year.
- **Comparison**: Alternative reading program or business-as-usual instructional approach.
- **Outcomes**: Quantitative measures of reading performance not created by program developers or study authors.
- **Study Designs**: Randomized experiments or quasi-experiments with matched comparisons.

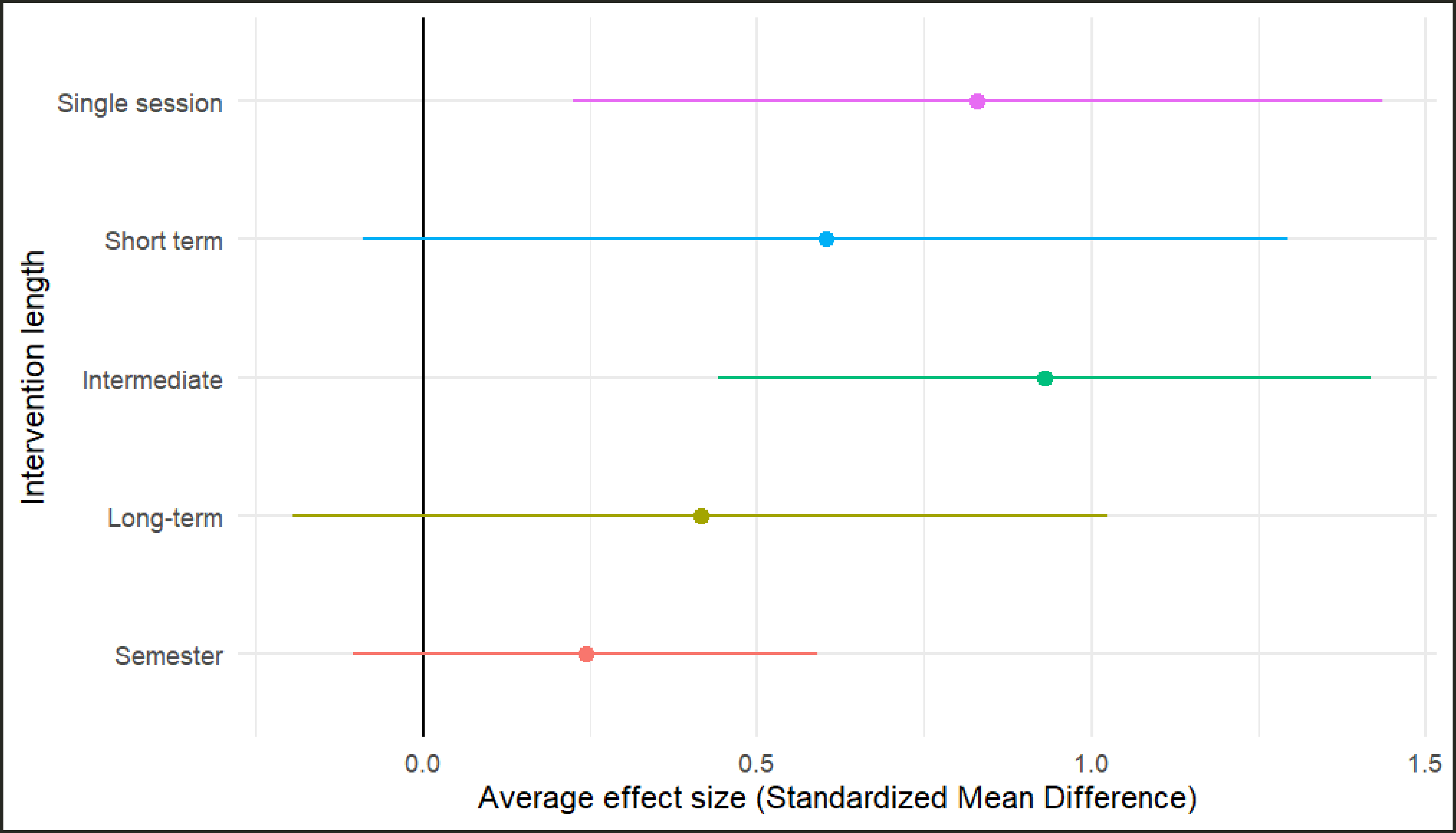
Tools

- Table 1: Summary of study characteristics
- Supplementary Table S1: Listing of included studies, their features, characteristics, and quality

Challenges

- Methods are available for meta-analysis of dependent effects, ability to analyze ***all*** relevant effect size estimates in a single, coherent model.
- In studies with many outcomes, need better guidelines about which to include.

Moderator analysis is correlational



What works for whom and under what conditions?

- **Moderator analysis** is used to investigate variation in effect sizes.
  - How are study characteristics related to magnitude of intervention effects?
  - Often a key goal for informing evidence-based practice.
- Moderator analysis through meta-regression:

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

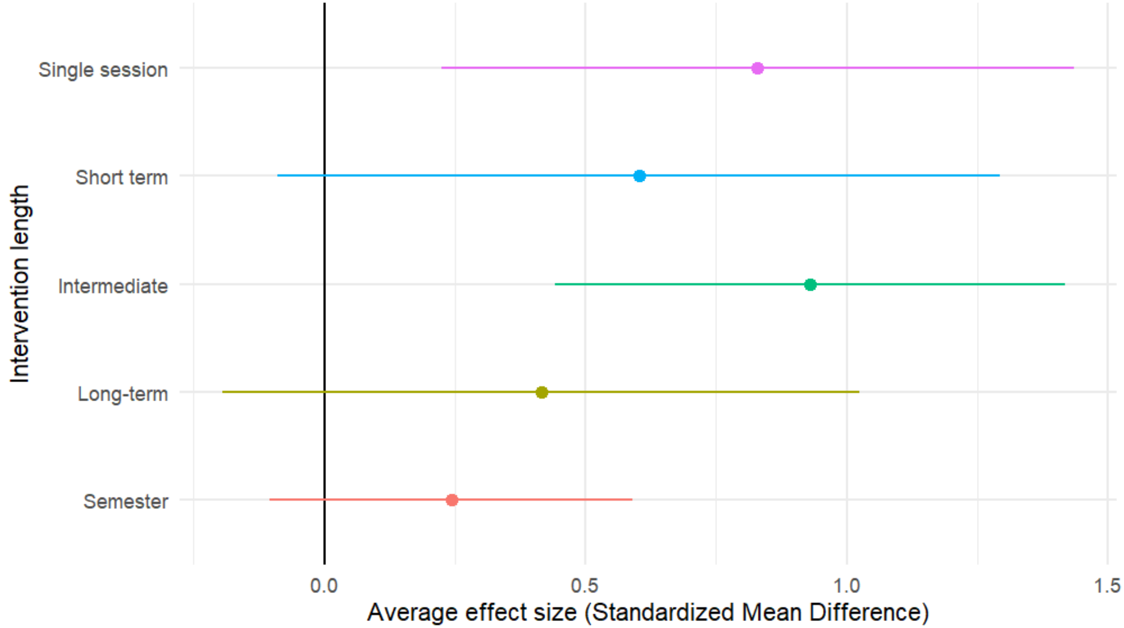
  - β_1 measures differences in average effects for studies differing on X_1.
  - Or for a categorical predictor with levels A, B, C, D:

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

- ***Moderating relations are not necessarily causal effects!***
  - Studies that differ on X_1 might differ in other ways too.

Is shorter better?

- Average effects of epistemic cognition interventions were **smaller** with longer-duration interventions:



- This does not necessarily imply that we should focus on shorter-duration interventions.

Number of effect sizes by intervention length and student level

Student level	Single session	Short term	Intermediate	Long-term	Semester
Elementary school students	0	0	1	5	0
Middle school students	0	0	3	0	2
High school students	6	5	4	0	3
College students	6	2	6	5	9
Graduate students	0	0	0	0	1
Other (like adults)	1	0	0	0	0

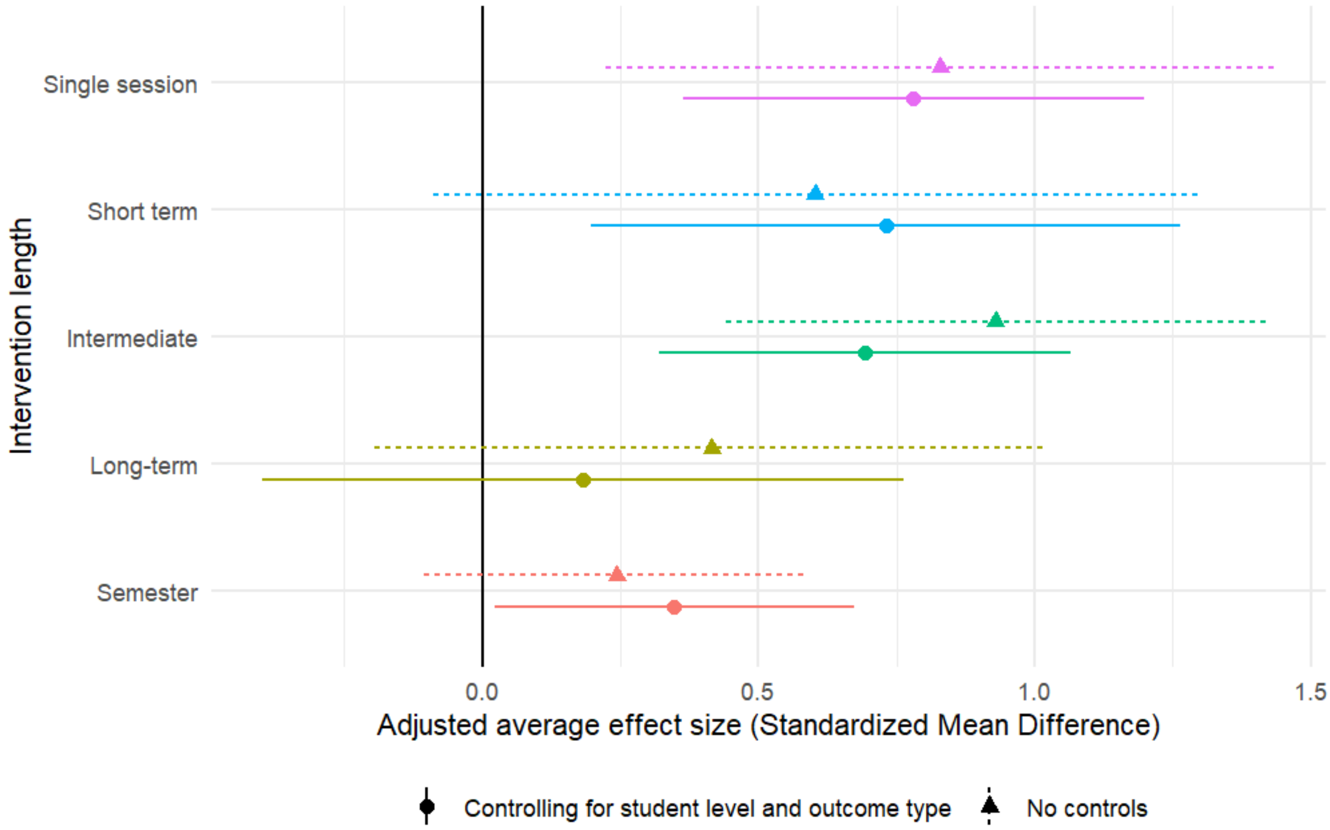
Number of effect sizes by intervention length and outcome measure

Outcome measure	Single session	Short term	Intermediate	Long-term	Semester
Researcher developed	13	7	14	5	6
Teacher developed	0	0	0	4	5
Standardized test	0	0	0	1	3
GPA (official)	0	0	0	0	1

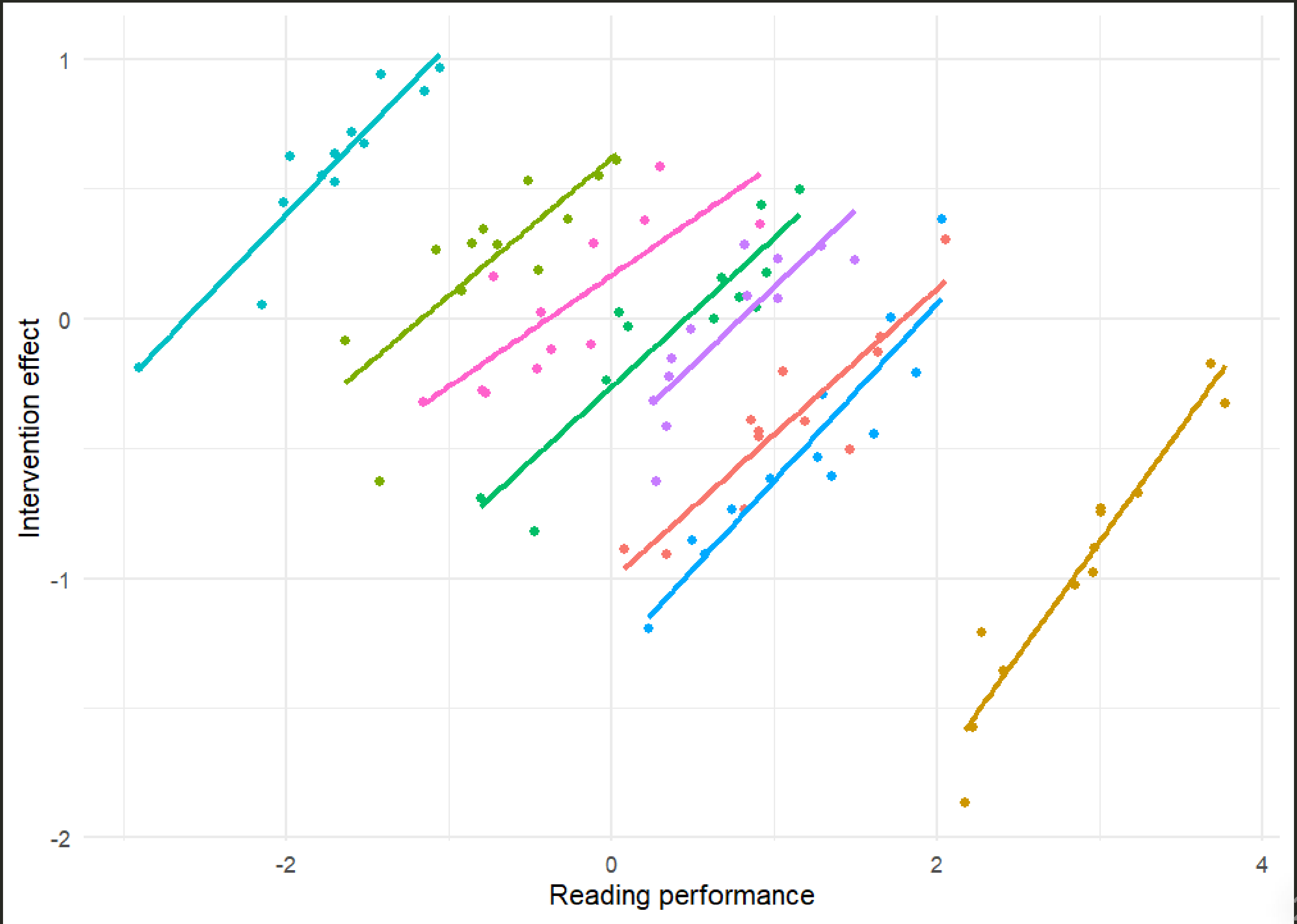
Tools

- Meta-regression, controlling for potential confounding variables:

$$T_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \epsilon$$



Aggregated data has limits



Conventional meta-analysis uses summary results

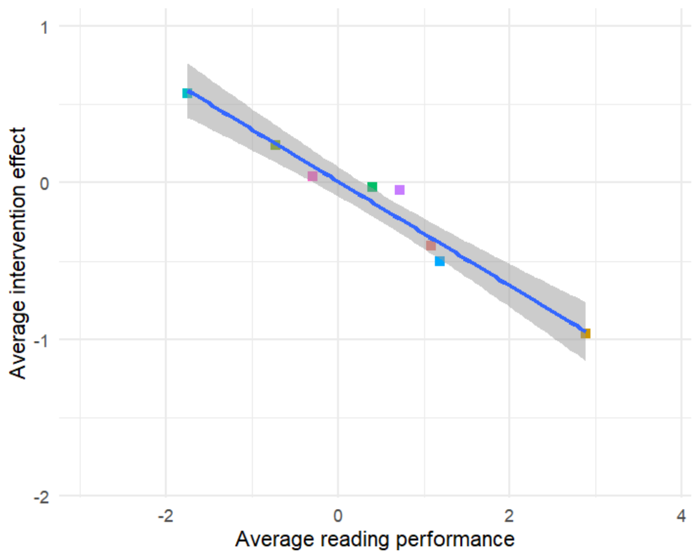
- Effect size estimates in meta-analysis are usually **summaries** or **averages** across individual participants.
    - Each T_i is an average intervention effect across participants in study i.
    - Studies differ in sample composition.
- We could use moderator analysis to understand how intervention effects vary in relation to sample characteristics:
    - Predicting average intervention effects based on initial academic performance:

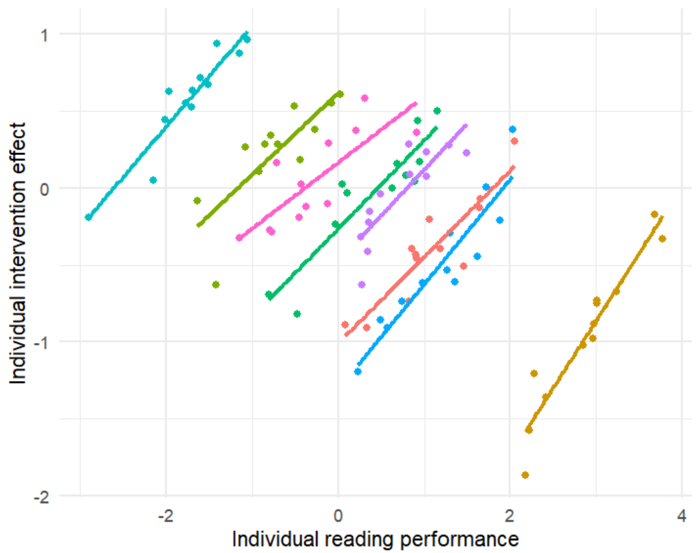
$$T_i = \beta_0 + \beta_1 (\text{Average baseline performance})_i + \epsilon_i$$

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

A stylized example

- Cheung and colleagues found that Success for All has larger effects for participant samples that had initially lower levels of reading performance.





- **Within-study relation differs from aggregate-level relation.**

Tools

- Controlling for contextual confounds
- Synthesis of **individual participant data**

Challenges

- Individual participant data is challenging to obtain and harmonize.
- Meta-analyses should make adjustments for differences in sample composition.
  - Targeted generalization, using tools from causal inference.

Four things to remember about meta-analysis

- Meta-analyses are about distributions.
- Inclusion criteria need to strike a balance.
- Moderator analysis is correlational.
- Aggregated data has limits.

Despite the limitations of meta-analysis (and moderator analyses in particular), research synthesis is still a critical tool for building evidence and informing practice/policy.