A Primer on Meta-Analysis

NSF ECR Hub PI Meeting 2026

Elizabeth Tipton
Department of Statistics and Data Science, Northwestern University

2026-07-01

Where we are

Terri covered the **systematic review**: framing the question, searching, screening, and coding studies and moderators.

That produces a coded dataset. This part asks:

- How do we put study outcomes on a **common scale**? (effect sizes)
- How do we summarize the **distribution** of the effects?
- How do we **explain** variation in effects?

A review organizes evidence. A meta-analysis quantifies it.

Effect sizes

The standardized mean difference

Studies report outcomes on different scales.

- In order to make them comparable, we rescale each into a common metric.
- For two-group comparisons, the workhorse is the **SMD**.

Population quantity:

$$\delta = \frac{\mu_T - \mu_C}{\sigma}$$

The SMD is on the **standard deviation** scale.

Interpretation of the SMD

The SMD is on the **standard deviation** scale. *e.g., if* $\delta = 0.2$ *we might say that, on average, the intervention increased outcomes 0.20 standard deviations.*

Keep in mind that for the same intervention:

- SMDs tend to be larger for narrow, proximal outcomes than broad, distal outcomes
- SMDs tend to be smaller in RCTs than in QEDs
- SMDs tend to be larger for more intensive interventions than light touch ones
- And so on …

Estimation of the SMD

The sample estimate (Cohen's d) uses a pooled SD to estimate δ:

$$d = \frac{\bar{Y}_T - \bar{Y}_C}{s_p}, \qquad s_p = \sqrt{\frac{(n_T - 1)s_T^2 + (n_C - 1)s_C^2}{n_T + n_C - 2}}$$

But d is upward-biased in small samples ($n < 30$). The correction is a multiplier J that depends only on $df = n_T + n_C - 2$:

$$g = J \cdot d, \qquad J \approx 1 - \frac{3}{4\,df - 1}$$

This is **Hedges g.**

Variance of g

The (approximate) sampling variance of g:

$$v_g = \frac{n_T + n_C}{n_T\, n_C} + \frac{g^2}{2(n_T + n_C)}$$

Each study contributes a pair $(g_i,\ v_i)$. Everything downstream is built from these pairs.

Other effect sizes

While the SMD is the most common effect size in education research, there are other available effect size families:

- for $X = 0, 1$ and Y contiuous: the **SMD**-family
- for X and Y continuous: the **correlation** family
- for $X = 0, 1$ and $Y = 0, 1$: the **log odds** family

These scales are not on the same metric, but can be converted to one another using a variety of formulas.

Combining effect sizes

Two models

There are two general models used in meta-analysis:

Fixed-effect (common-effect): all studies estimate the *same* θ.

$$g_i = \theta + e_i, \qquad e_i \sim N(0, v_i)$$

Random-effects: studies estimate *related but different* true effects.

$$g_i = \theta + u_i + e_i, \qquad u_i \sim N(0, \tau^2), \qquad e_i \sim N(0, v_i)$$

τ^2 is the **between-study variance**. In education $\tau^2 > 0$ is the rule, so **random effects is the sensible default.**

The random-effects summary

Recall that we have pairs (g_i, v_i) where the v_i differ across studies. This means that the data is **heteroscedastic**. The solution is to use **inverse variance** weights.

The distribution of effects can be summarized in terms of its mean, θ and variance, τ^2.

The average θ is estimated using inverse-variance weights:

$$\hat{\theta} = \frac{\sum_i w_i \, g_i}{\sum_i w_i}, \qquad w_i = \frac{1}{v_i + \hat{\tau}^2}$$

The variance τ^2 is also estimated (REML is standard).

Describing the variation, not just the mean

It is important to report the full distribution of effect sizes.

The **prediction interval** describes where the range of true effects. Typically, we do so assuming that true effects are normally distributed across studies:

$$\hat{\theta} \pm t_{1-\alpha,k-2} \sqrt{\hat{\tau}^2 + \mathrm{SE}(\hat{\theta})^2}$$

For example, you might report 95-percent intervals: *The mean of the distribution is estimated as* $\hat{\theta} = 0.21$*, and* 95*-percent of true effect sizes are estimated to be between* 0.014 *and* 0.406*.*

Prediction intervals vs Confidence intervals

It is easy to get PIs and CIs confused.

- Keep in mind that the CI for $\hat{\theta}$ shrinks as k grows — it describes uncertainty about the *mean*.
- But the PI has to do with the distribution of true effects. As k grows the PI does not necessarily change.

Dependence: multiple effect sizes per study

The problem

Real reviews rarely give one effect size per study. A study may report:

- several outcomes (reading *and* math),
- multiple time points,
- several treatment arms vs. a shared control.

These effect sizes are **correlated**. Treating them as independent **understates uncertainty** and double-counts information.

A three-level view

Split the variation into between-study and within-study (between effect sizes inside a study):

$$g_{ij} = \theta + u_j + r_{ij} + e_{ij}$$

- $u_j \sim N(0, \tau^2)$ — between studies
- $r_{ij} \sim N(0, \omega^2)$ — between effect sizes within a study
- $e_{ij} \sim N(0, v_{ij})$ — sampling error

The total true variation is now $\tau^2 + \omega^2$, which is also used in the prediction interval.

Robust variance estimation (RVE)

We rarely know the within-study correlations. **RVE** corrects the standard errors *empirically* instead of requiring us to specify them correctly.

- Fit a working model (the three-level model above, with an assumed correlation ρ).
- Compute **cluster-robust** standard errors, clustering by study.
- Apply small-sample corrections; refer tests to t with estimated df.

A working model plus RVE — the **CHE-RVE** approach — is the current practical default for dependent effects, and is built into `metafor`.

What RVE buys you

- Valid inference even when the assumed correlation ρ is wrong.
- Correct standard errors that respect clustering by study.

The cost: with few clusters (studies), the estimated df is low and intervals can be very wide.

Importantly, this means that when you report results, you need to report: $(\hat{\theta}, SE(\hat{\theta}), df, p - value)$.

Moderators

Meta-regression

When $\tau^2 + \omega^2 > 0$, and especially when the PI is wide, the purpose you need to ask *why* effects differ.

Moderators (the covariates coded earlier) enter as predictors:

$$g_{ij} = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + u_i + r_{ij} + e_{ij}$$

- Interpreted as in regression, on the effect-size scale.
- Goal: does one or more moderators account for part of $\tau^2 + \omega^2$?

Importantly, you will continue to use RVE here when effect sizes are dependent.The degrees of freedom vary by coefficient, and can be much smaller than the number of studies.

Cautions

It is important to be very careful in interpreting moderator relationships:

- Moderators are *observational* even when studies are RCTs (associations confounded across studies, not causal). *This means controlling for other variables is useful.*
- Power is typically very low (the max df = k). *Lack of significance does not mean that the null hypothesis is true.*
- The model that explains the most of the variation τ^2 is a predictive model. *This means that the absence of a variable in the model might be because it is highly correlated with other variables, not that it doesn't matter.*

Wrapping up

The pipeline, end to end

1. Code each study into an effect size (g_i, v_i).
2. Fit a random-effects model: estimate θ *and* τ^2.
3. Report the distribution: provide a prediction interval.
4. Handle multiple effect sizes per study with a multilevel model + RVE.
5. Probe moderators with meta-regression — cautiously.

The average is the start of the analysis, not the end.

Tools: `metafor` (Viechtbauer) does all of the above — estimation, RVE, prediction intervals, and plots.

Appendix: doing it all in metafor

```r
library(metafor)

# 1. Effect sizes: Hedges g (yi) and variances (vi)
dat <- escalc(measure = "SMD",
              m1i = mT, sd1i = sdT, n1i = nT,
              m2i = mC, sd2i = sdC, n2i = nC,
              data = raw)

# 2. Single effect size per study: random-effects model
res <- rma(yi, vi, data = dat)
predict(res)            # CI and prediction interval

# 3. Multiple effect sizes per study (CHE working model):
#    build the sampling var-cov matrix assuming within-study r = 0.6
V <- vcalc(vi, cluster = study, obs = outcome, rho = 0.6, data = dat)

che <- rma.mv(yi, V,
              random = ~ 1 | study/outcome,
              data = dat)

# 4. RVE inference, clustering by study (CR2 + Satterthwaite df)
robust(che, cluster = dat$study, clubSandwich = TRUE)

# 5. Meta-regression with a moderator, same RVE step
mod <- rma.mv(yi, V, mods = ~ grade_level,
```