Model-Building Considerations in Meta-Analysis of Dependent Effect Sizes

James E. Pustejovsky

May 17, 2024

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

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

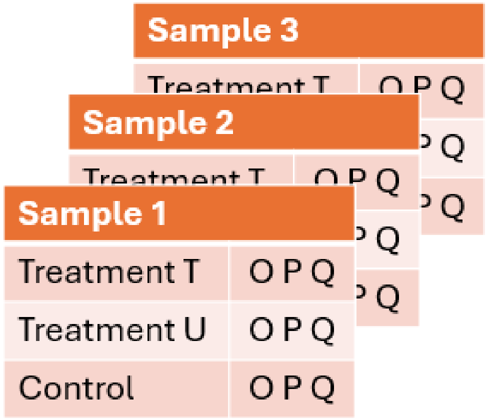
→ d_{12}, d_{22}, d_{32}

Multiple treatment conditions compared to a common control

Treatment T	O
Treatment U	O
Control	O

→ d_{T3}, d_{U3}

Multiple samples/sites within a study



Multiple specific groups within a sample

Group A	Treatment	O Q
	Control	O Q
Group B	Treatment	O Q
	Control	O Q

Tanner-Smith & Lipsey (2015). Brief alcohol interventions for adolescents and young adults: A systematic review and meta-analysis.

- 185 studies, 1446 effect size estimates
    - Standardized mean differences comparing alcohol consumption outcomes of intervention participants to comparison participants.
    - Multiple outcome measures
    - Multiple follow-up times
    - Multiple treatment conditions
    - Multiple comparison groups
    - 1-108 effect size estimates per study (median = 6, IQR = 3-12)

Chen et al. (2020). Gender Differences in Life Satisfaction Among Children and Adolescents: A Meta-Analysis.

- 101 effect size estimates drawn from 52 samples in 46 studies.
    - Standardized mean differences comparing boys versus girls on life satisfaction self-report measures.
    - Multiple distinct samples nested within studies.
    - Multiple measures of life satisfaction collected on same sample.

Notation

- K effect sizes
- J samples/experiments
- Sample j includes k_j (possibly dependent) effect size estimates
- Effect size i in study j is an estimate of parameter θ_{ij}
- Effect size i in study j has estimate T_{ij} with sampling variance $V_{ii,j}$, plus predictors $\mathbf{x}_{ij} = \left(x_{1ij}, \ldots, x_{pij}\right)$.
- Covariance between effect size estimates h and i in study j is

$$\text{Cov}(T_{hj}, T_{ij} | \theta_{hj}, \theta_{ij}) = V_{hi,j}$$

$$\begin{bmatrix} T_{1j} \\ T_{2j} \\ \vdots \\ T_{k_j j} \end{bmatrix} = \begin{bmatrix} \theta_{1j} \\ \theta_{2j} \\ \vdots \\ \theta_{k_j j} \end{bmatrix} + \begin{bmatrix} e_{1j} \\ e_{2j} \\ \vdots \\ e_{k_j j} \end{bmatrix} \qquad \text{where} \qquad \text{Var}\left(\begin{bmatrix} e_{1j} \\ e_{2j} \\ \vdots \\ e_{k_j j} \end{bmatrix}\right) = \begin{bmatrix} V_{11,j} & V_{12,j} & \cdots & V_{1k_j,j} \\ V_{21,j} & V_{22,j} & \cdots & V_{2k_j,j} \\ \vdots & \vdots & \ddots & \vdots \\ V_{k_j 1,j} & V_{k_j 2,j} & \cdots & V_{k_j k_j,j} \end{bmatrix}$$

Working model

$$T_{ij} = \underbrace{\mathbf{x}_{ij}\boldsymbol{\beta}}_{\text{fixed predictors}} + \underbrace{u_{ij} + e_{ij}}_{\text{error structure}}$$

- A *tentative* model for the error structure, which might be only a *rough approximation* to the true data-generating process.

Building a working model

1. ***Estimate or make assumptions*** about covariances between sampling errors.

2. Model the structure of the true effects, often using a ***multivariate*** or ***multilevel*** model (allowing for within-sample heterogeneity).

3. Use cluster-robust variance estimation methods to protect against mis-specification.

Methods for handling dependent effect sizes

Becker (2000) describes four broad strategies for handling dependent effects.

Ignore the dependence

- Not usually advisable

Combine estimates

- Aggregate (average) dependent effect sizes to the sample level.

Sub-classify effects (shifting unit-of-analysis)

- Create multiple subsets of effect sizes.
- Aggregate within subsets so that each sample has at most one effect size estimate per subset.

Model the dependence

- Multivariate meta-analysis
- Multilevel meta-analysis
- Working models with robust variance estimation

Aggregating effect size estimates

- Take a simple or weighted average of effect sizes from each study.

$$\bar{T}_j = \frac{1}{w_{\bullet j}} \sum_{i=1}^{k_j} w_{ij} T_{ij}, \qquad w_{\bullet j} = \sum_{i=1}^{k_j} w_{ij}$$

- The variance of $\bar{T}_j$ depends on $V_{ii,j}$ and on the *covariances* $V_{hi,j}$:

$$V_{\bullet j} = \mathrm{Var}\left(\bar{T}_j\right) = \frac{1}{w_{\bullet j}^2} \sum_{h=1}^{k_j} \sum_{i=1}^{k_j} w_{hj} w_{ij} V_{hi,j}$$

- If covariances are unknown or hard to calculate, we might assume that there is a **constant sampling correlation**, ρ among the effect size estimates $\left(V_{hi,j} = \rho \sqrt{V_{hh,j} \times V_{ii,j}}\right)$.

Gender differences in life satisfaction

```r
Chen_agg <- aggregate(
  x = Chen,
  cluster = SampleID,
  obs = EffectID,
  rho = 0.5
)

Chen_agg_RE <- rma.uni(
  yi = yi, vi = vi,
  data = Chen_agg,
  method = "REML"
)
```

```
##
## Random-Effects Model (k = 52; tau^2 estimator: REML)
##
## tau^2 (estimated amount of total heterogeneity): 0.0275 (SE = 0.0065)
## tau (square root of estimated tau^2 value):      0.1659
## I^2 (total heterogeneity / total variability):   87.99%
## H^2 (total variability / sampling variability):  8.33
##
## Test for Heterogeneity:
## Q(df = 51) = 429.4501, p-val < .0001
##
## Model Results:
##
## estimate      se    zval    pval    ci.lb   ci.ub
##   0.0255  0.0253  1.0055  0.3147  -0.0242  0.0751
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Meta-analysis of aggregated data

- Random effects meta-analysis for the aggregated effect sizes:

$$\bar{T}_j = \bar{\mathbf{x}}_j \boldsymbol{\beta} + u_j + \bar{e}_j, \qquad \text{Var}(\bar{e}_j) = V_{\bullet j}$$

- Pustejovsky & Chen (2024) show that this is ***exactly equivalent*** to a model for the raw effect sizes:

$$T_{ij} = \bar{\mathbf{x}}_j \boldsymbol{\beta} + u_j + e_{ij}$$

  where $\text{Var}(e_{ij}) = V_{ii,j}, \text{Cov}(e_{hj}, e_{ij}) = V_{hi,j}$.

- Pustejovsky & Tipton (2022) call this a "correlated effects" model.

Gender differences in life satisfaction

```
Vmat <- vcalc(
  data = Chen,sparse = TRUE,
  vi = vi,
  cluster = SampleID, obs = EffectID,
  rho = 0.5
)

Chen_CE <- rma.mv(
  yi = yi, V = Vmat,
  random = ~ 1 | SampleID,
  data = Chen,
  method = "REML", sparse = TRUE
)
```

Model	K	Average ES	SE (model)	SE (robust)	Heterogeneity SD	Q statistic
Aggregated random effects	52	0.025	0.025	0.028	0.166	429.450
Correlated effects	101	0.025	0.025	0.028	0.166	813.439

- Point estimates, SEs, confidence intervals are identical.
- Robust SEs and confidence intervals are identical.
- Q statistics differ
  - Aggregated effects Q measures excess heterogeneity ***of averaged effect size estimates***
  - Correlated effects Q measures excess heterogeneity ***of raw effect size estimates***

Aggregating effect sizes is the same as fitting a working model with between-sample heterogeneity (but no within-sample heterogeneity).

So what?

- Useful heuristic: excluding random effects is just like aggregating.
- Using the multivariate representation allows for comparison to other multivariate working models.
- If the correlated effects model is justified, then aggregating is justified.
  - Computational shortcut.
  - Figures/graphical diagnostics can use aggregated effect size estimates.

Sub-classifying/Shifting unit-of-analysis

- Classify effect sizes into categories where each study contributes ≤ 1 effect size per category.
- If there are still multiple effect sizes from the same study within a given category, aggregate them together (Cooper, 1998).
- Run meta-analysis **separately** for each category.

LS	yi	sei
Overall	-0.312	0.128
Overall	0.108	0.066
Overall	0.150	0.092
Friends	-0.055	0.092
Living environment	0.226	0.092
Family	0.050	0.092
Self	0.095	0.092
Overall	0.033	0.053
School	-0.547	0.177
Self	0.643	0.178
Overall	-0.142	0.074
Overall	-0.034	0.047
Living environment	0.043	0.066
Family	-0.110	0.066
Friends	-0.398	0.067
School	-0.382	0.067
Self	-0.089	0.066
Overall	-0.080	0.053
School	-0.293	0.049
School	-0.181	0.050

$\longrightarrow$

LS	yi	sei
Overall	-0.141	0.129
Overall	-0.169	0.130
Overall	0.121	0.104
Overall	-0.312	0.128
Overall	0.108	0.066
Overall	0.150	0.092
Overall	0.033	0.053
Overall	-0.142	0.074
Overall	0.000	0.057
Overall	0.000	0.057
Overall	0.000	0.057
Overall	0.000	0.057
Overall	0.000	0.057
Overall	-0.034	0.047
Overall	-0.080	0.053
Overall	0.374	0.040
Overall	0.191	0.087
Overall	0.101	0.056
Overall	0.076	0.046
Overall	0.116	0.047

LS	yi	sei
School	-0.547	0.177
School	-0.382	0.067
School	-0.293	0.049
School	-0.181	0.050
School	0.039	0.137
School	-0.411	0.123
School	-0.490	0.117
School	-0.442	0.153
School	-0.335	0.105
School	0.065	0.096
School	-0.154	0.043
School	-0.264	0.115
School	-0.128	0.050
School	-0.328	0.076
School	-0.123	0.041
School	-0.121	0.103

LS	yi	sei
Self	0.095	0.092
Self	0.643	0.178
Self	-0.089	0.066
Self	0.368	0.138
Self	0.302	0.123
Self	-0.164	0.115
Self	-0.182	0.152
Self	0.041	0.104
Self	0.421	0.097
Self	0.097	0.054
Self	0.164	0.115
Self	0.145	0.041
Self	0.416	0.104

LS	yi	sei
Family	0.050	0.092
Family	-0.110	0.066
Family	0.164	0.137
Family	0.188	0.122
Family	-0.133	0.115
Family	-0.210	0.152
Family	-0.137	0.104
Family	0.063	0.096
Family	-0.032	0.115
Family	0.201	0.161
Family	0.000	0.041
Family	-0.146	0.103

LS	yi	sei
Living environment	0.226	0.092
Living environment	0.043	0.066
Living environment	0.396	0.123
Living environment	-0.182	0.115
Living environment	-0.361	0.153
Living environment	-0.307	0.105
Living environment	0.306	0.096
Living environment	-0.079	0.115
Living environment	0.013	0.041
Living environment	0.133	0.103

LS	yi	sei
Friends	-0.055	0.092
Friends	-0.398	0.067
Friends	0.376	0.138
Friends	-0.363	0.123
Friends	-0.208	0.115
Friends	-0.115	0.152
Friends	-0.121	0.104
Friends	-0.013	0.096
Friends	-0.096	0.115
Friends	-0.089	0.041
Friends	-0.089	0.103

Gender differences by life satisfaction domain

```
Chen_overall <- rma.uni(yi = yi, vi = vi, data = Chen, subset = LS == "Overall")
Chen_school <- rma.uni(yi = yi, vi = vi, data = Chen, subset = LS == "School")
Chen_self <- rma.uni(yi = yi, vi = vi, data = Chen, subset = LS == "Self")
```

LS	K	Average ES	SE (model)	SE (robust)	Heterogeneity SD	Q statistic
Overall	39	0.069	0.025	0.028	0.140	234.607
School	16	-0.237	0.038	0.030	0.123	49.526
Self	13	0.162	0.062	0.062	0.195	53.073
Family	12	-0.021	0.029	0.030	0.038	15.297
Living environment	10	0.026	0.075	0.084	0.215	43.946
Friends	11	-0.116	0.057	0.055	0.156	37.028

Sub-classifying/Shifting unit-of-analysis

- Let $T_{cj}, V_{icj}, \mathbf{x}_{cj}$ correspond to effect i in category c in study j.
- The model for studies in sub-class c:

$$T_{cj} = \mathbf{x}_{cj}\boldsymbol{\beta}_c + u_{cj} + e_{cj}$$

  where $\text{Var}(u_{cj}) = \tau_c^2$ and $\text{Var}(e_{cj}) = V_{cj}$.

- Pustejovsky & Chen (2024) show that meta-analysis of sub-classes is exactly equivalent to a model ***for the full data*** that assumes:

$$T_{cj} = \underbrace{\mathbf{x}_{cj}\boldsymbol{\beta}_c}_{\mathbf{x}\times\text{category interactions}} + \underbrace{u_{cj}}_{\text{separate random effects}} + \underbrace{e_{cj}}_{\text{independent sampling errors}}$$

  - Distinct β coefficients for each sub-class
  - Effect size estimates from different sub-class are independent
  - Heterogeneity differs by sub-class

- Pustejovsky & Tipton (2022) call this a "subgroup correlated effects" model.

Subgroup correlated effects model

```
Chen_LS <- rma.uni(
  yi = yi, vi = vi,
  data = Chen,
  mods = ~ 0 + LS,
  scale = ~ 0 + LS, link = "identity",
)

clubSandwich::conf_int(
  Chen_LS, vcov = "CR2", cluster = Chen$StudyID
)
```

```
V_sub <- vcalc(
  data = Chen, sparse = TRUE,
  vi = vi,
  cluster = SampleID, obs = EffectID,
  rho = 0.5,
  subgroup = LS
)

Chen_SCE <- rma.mv(
  yi = yi, V = V_sub,
  data = Chen,
  mods = ~ 0 + LS,
  random = ~ LS | SampleID, struct = "DIAG",
) |>
  robust(cluster = StudyID, clubSandwich = TRUE)
```

Sub-classifying effect sizes (shifting the unit of analysis) is the same as fitting a multivariate model that treats effects in different sub-classes as independent.

So what?

- Useful heuristic: sub-classifying is just like a model that treats different sub-classes as independent.
- Working model representation allows comparisons of different sub-classes because they're all represented in a single model.
- Results based on sub-classifying are a useful point of comparison to results based on working models that include ***cross-category dependence***

Discussion

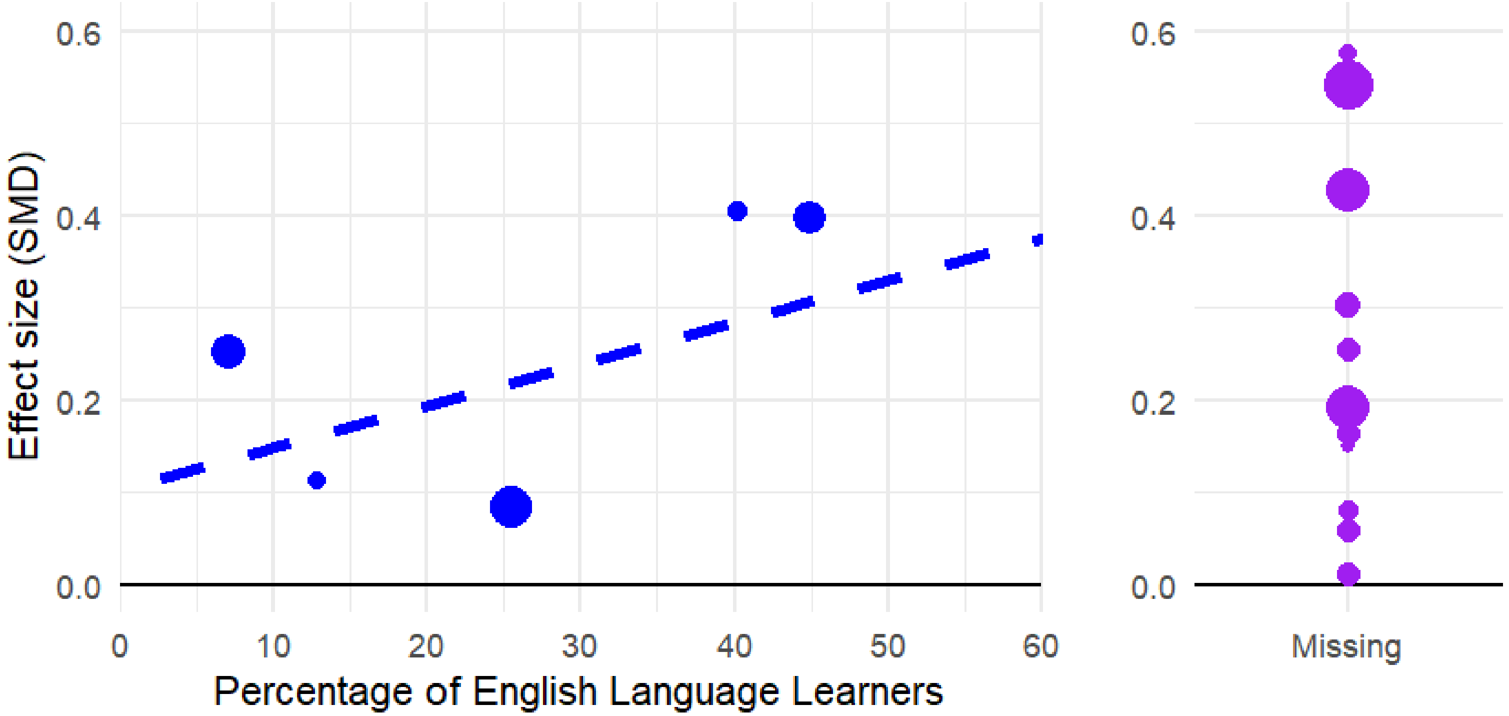
- Older "ad hoc" methods for handling dependent effect sizes are equivalent to multivariate working models.
- Multivariate working model representations are useful for model comparison, critique, and sensitivity analysis.
- Equivalence relationships provide helpful heuristics for constructing working models.
- Robust variance estimation is very often helpful, with any of these working models (even ad hoc methods).

Equity-related moderator analysis

- In syntheses of educational intervention studies, our goal is to understand the **distribution of program impacts**.
  - Equity-related moderator analyses seek to address questions of **who benefits** from an intervention and **how benefits and harms are distributed** across students.
- Moderator analyses examine variation in effect size based on characteristics of primary study participants and contexts:
  - Participants' family income level
  - Participant racial/ethnic groups
  - Participant English Language Learner status
  - School urbanicity

Synthesis of study-level average effects

- Traditional synthesis involves examining associations between average effect sizes and aggregate sample characteristics.



Source: Neitzel, et al. (2020). Data archive for "Success for All: A Quantitative Synthesis of U. S. Evaluations." Towson, MD: Center for Research and Reform in Education (CRRE), Johns Hopkins University. Retrieved from https://github.com/aj-neitzel/Success-for-All-A-Quantitative-Synthesis-of-U.-S.-Evaluations

Synthesis of dependent effect sizes

- Results on ***multiple outcome measures***
- Results at ***multiple follow-up times***
- Results for each of ***several subgroups***
- Results from each of ***multiple samples*** or ***multiple specific groups***

Study	Sample	Subgroup	Followup	ELL %	N	ES 1	ES 2	ES 3
A	A.1	Non-ELL	Short	0	108	0.05	0.26	0.16
A	A.1	ELL	Short	100	21	0.13	-0.23	0.15
B	B.1	Non-ELL	Short	0	48	0.36	-0.03	0.11
B	B.1	ELL	Short	100	36	0.45	0.11	-0.07
B	B.1	Non-ELL	Long	0	48	-0.07	0.86	0.03
B	B.1	ELL	Long	100	36	1.06	0.42	0.22
C	C.1	Mix	Short	15	77	-0.30	0.23	0.05
C	C.2	Mix	Short	22	46	-0.29	0.07	0.53
C	C.3	Mix	Short	12	52	0.20	0.12	0.17
D	D.1	Mix	Short	36	114	-0.05	0.31	0.46
D	D.1	Mix	Long	36	114	-0.23	0.20	0.40
D	D.2	Mix	Short	31	97	-0.14	0.54	0.46
D	D.2	Mix	Long	31	97	0.05	0.66	0.21

Direct evidence

- Reported effect size estimates for each of multiple subgroups.
- Provides estimates of ***individual-level variation*** in impacts.
- Study-level operational features are held constant.

Study	Followup	ELL %	N	ES 1	ES 2	ES 3
A	Short	0	108	0.05	0.26	0.16
A	Short	100	21	0.13	-0.23	0.15
B	Short	0	48	0.36	-0.03	0.11
B	Short	100	36	0.45	0.11	-0.07
B	Long	0	48	-0.07	0.86	0.03
B	Long	100	36	1.06	0.42	0.22

Contextual evidence

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

Study	Sample	Followup	N	ELL %	ES 1	ES 2	ES 3
A	A.1	Short	129	16.28	0.06	0.18	0
B	B.1	Long	84	42.86	0.42	0.67	0
B	B.1	Short	84	42.86	0.40	0.03	0
C	C.1	Short	77	15.00	-0.30	0.23	0
C	C.2	Short	46	22.00	-0.29	0.07	1
C	C.3	Short	52	12.00	0.20	0.12	0
D	D.1	Long	114	36.00	-0.23	0.20	0
D	D.1	Short	114	36.00	-0.05	0.31	0
D	D.2	Long	97	31.00	0.05	0.66	0
D	D.2	Short	97	31.00	-0.14	0.54	0

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

...and should be analyzed as such.

- Meta-analyze the direct evidence (subgroup-specific effect sizes) alone, excluding the contextual evidence.

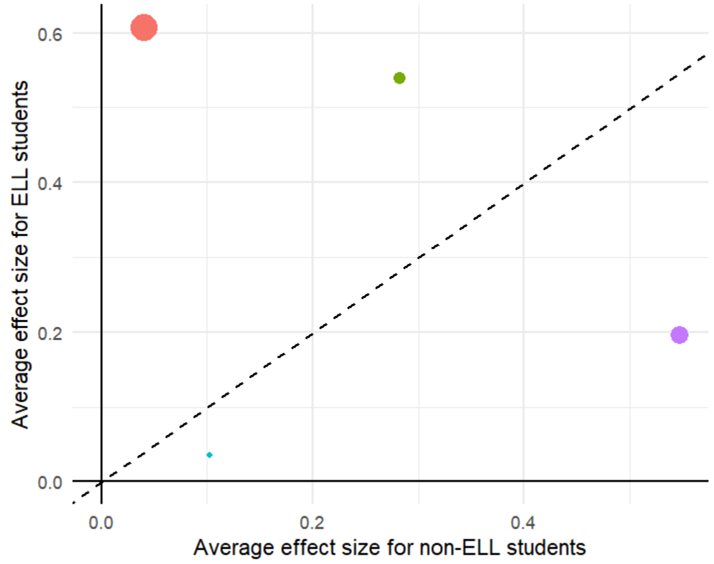
and/or

- 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 (subgroup-specific effect sizes) in a separate meta-analysis, excluding the contextual evidence.

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



Center by sample

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

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

- Estimate a meta-regression with sample-centered and sample-aggregate predictors:

$$ES_{ij} = \beta_0 + \beta_1 \underbrace{\left(ELL\%_{ij} - \overline{ELL\%}_j\right)}_{\text{direct evidence}} + \beta_2 \underbrace{\left(\overline{ELL\%}\right)_j}_{\substack{\text{contextual} \\ \text{evidence}}} + u_{ij} + e_{ij}$$

  - $\hat{\beta}_1$ is based only on samples providing direct evidence
  - $\hat{\beta}_2$ is based on sample-level aggregated effect sizes

Current practice

- We reviewed empirical meta-analysis projects funded by the Institute of Education Sciences between 2002 and 2018.
- 25 projects included "meta-analysis" in project description and had associated journal article reporting a meta-analysis.

Feature	Category	N	Pct
Any moderator analysis		24	96
Student characteristic moderators		16	64
Centering	Grand-mean	3	12
	Sample-mean	1	4
	Not specified	1	4
Working model	Correlated effects	9	36
	Aggregated effects	7	28
	Hierarchical effects	3	12
	Independent effects	2	8
	Multi-level	2	8

Further Recommendations

- Prior to conducting moderator analysis, **describe the structure of the evidence** on equity-related student characteristics.

Variable	Reported N ES (%)	Reported N Studies (%)	M	SD	Within-Study Variation N Studies (%)
Grade	1061 (96)	176 (92)	3.32	2.93	26 (14)
Male Pct	777 (70)	124 (65)	0.52	0.14	45 (32)
White Pct	656 (59)	109 (57)	0.40	0.27	41 (31)
Economic Disadvantage Pct	462 (42)	77 (40)	0.57	0.24	27 (28)
ELL Pct	385 (35)	56 (29)	0.22	0.24	23 (35)
SPED Pct	316 (28)	48 (25)	0.20	0.28	19 (33)

Source: Williams et al. (2022). Heterogeneity in Mathematics Intervention Effects: Evidence from a Meta-Analysis of 191 Randomized Experiments.

- If student characteristics are of focal interest, **use data extraction strategies to maximize amount of direct evidence.**

Limitations and future directions

- Data availability is a major limitation
  - Common to have missing information about sample-average characteristics.
  - Subgroup-specific results available only for a small subset of studies.
- Selective reporting of subgroup analysis could create biases in direct evidence (Hahn et al., 2000).
- Need to further develop working models for synthesizing direct and contextual evidence together.

Thanks for your attention!

James E. Pustejovsky
pustejovsky@wisc.edu
https://jepusto.com