Meta-Analysis Workshop

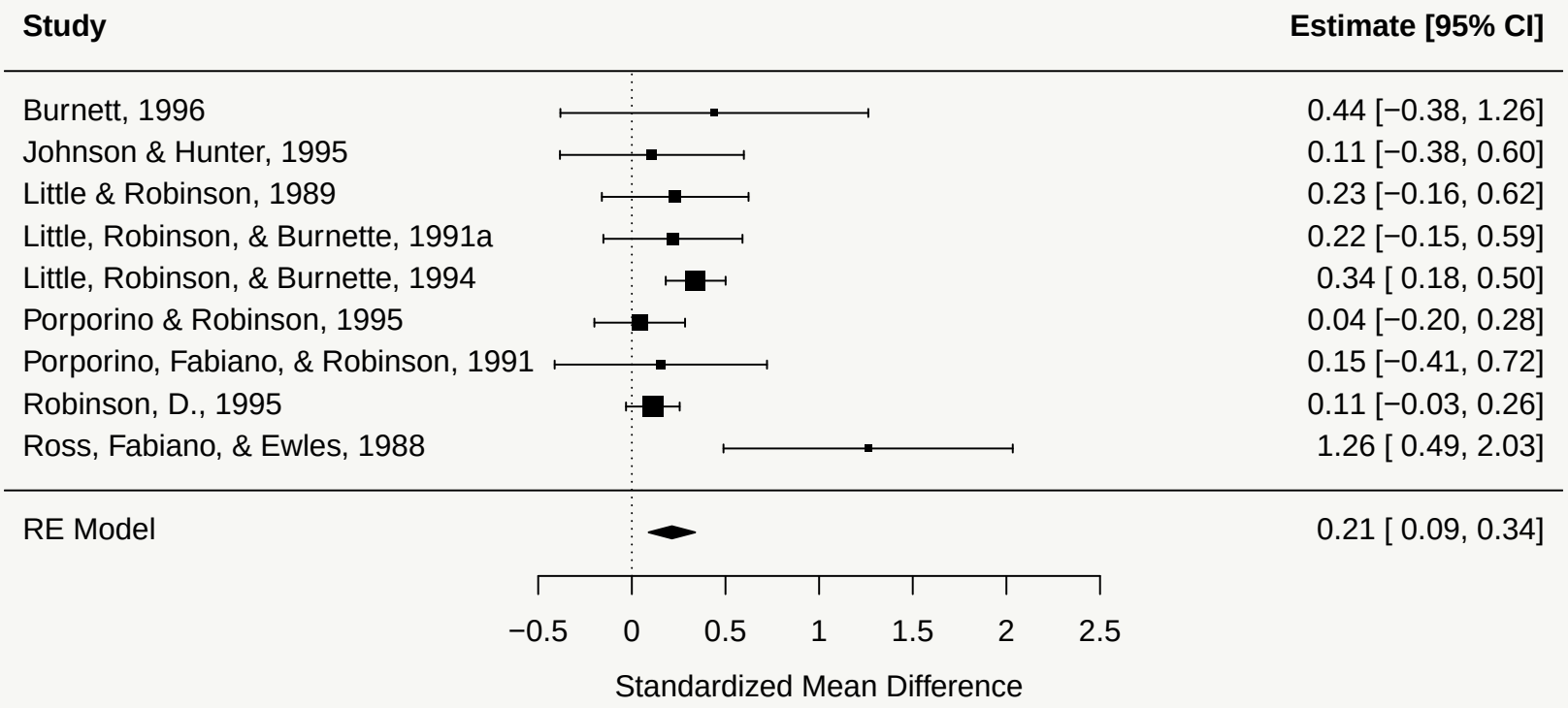
David B. Wilson, PhD

June 23-26, 2026

George Mason University
Department of Criminology, Law & Society
dwilsonb@gmu.edu

The End-Game

Forest-Plot of Standardized Mean Differences and 95% Confidence Intervals for the Effects of Cognitive-Behavioral Programs on Recidivism



Overview

Overview: Day 1

Day 1 Plan

- Historical background
- Logic of Meta-analysis
- Effect sizes
  - Common types
  - Computing standardized mean difference effect sizes
  - Computing odds ratio and risk ratio effect sizes
- Explain the effect size exercise
- Elements of a systematic review

Overview: Day 2

Day 2 Plan

- Review effect size exercise
- Basic meta-analysis methods
- Measure of heterogeneity
- Random-effects versus fixed-effect
- Prediction intervals
- Forest plots
- Explain the analysis exercise

Overview: Day 3

Day 3 Plan

- Review analysis exercise
- Moderator analysis
  - Subgroup analysis (analog to the ANOVA)
  - Meta-analytic regression
- Dependent effect sizes

Overview: Day 4

Day 4 Plan

- Review analysis exercise
- Publication bias
- Reporting and PRISMA 2020
- Wrap-up and Q&A

Historical Background

A Great Debate

- Hans Eysenck (1952): Claimed that psychotherapy does not work based on a narrative review
- A dizzying array of mixed results followed
- Gene Glass (with Mary Lee Smith) averaged effect sizes from 375 studies concluded that psychotherapy does work
- Gene Glass coined the term **meta-analysis**

Deep Roots

- Karl Pearson (1904): averaged correlations between inoculation for Typhoid fever and mortality
- R. A. Fisher (1944) noted that independent studies individually may not be significant, yet the aggregate may be improbable
- W. G. Cochran (1953) developed methods of averaging means across studies (basis of inverse variance weighted meta-analysis)
- A. Wicker (1967) averaged correlations between attitudes and behavior
- Concurrent with Smith and Glass (1977) were
  - Hunter and Schmidt (1977) *Validity generalization*
  - Rosenthal and Rubin (1978) *Interpersonal expectancy effects*

Logic of Meta-analysis

Logic of Meta-analysis

- Narrative review methods:
  - Focus on statistical significance
  - Lack of transparency and replicability
- Weakness of focusing on statistical significance:
  - A significant effect is a strong conclusion
  - A nonsignificant effect is a weak conclusion
  - How do you balance a collection of significant and non-significant effects? (Hint: you don't!)

Logic of Meta-analysis

- Meta-analysis:
  - Focuses on the **direction** and **magnitude** of effect
  - Approaches the task as a research endeavor
  - Examines the pattern of evidence across studies
    - average effect
    - consistency of effects
    - relationship between study features and effects

Research Suitable for Meta-analysis

Forms of Research Findings Suitable to Meta-analysis

- Central tendency (e.g., prevalence rates)
- Pre-post contrasts
- Group contrasts, naturally occurring (e.g., boys versus girls)
- Group contrasts, experiment (e.g., treatment versus control)

Forms of Research Findings Suitable to Meta-analysis

- Association between variables
  - Correlational studies (e.g., cross-sectional surveys)
  - Regression models (usually one coefficient is of interest)
  - Measurement (psychometric) research
  - Fully multivariate models (SEM, factor analysis)

Effect Sizes

The Effect Size: The Heart and Soul of Meta-analysis

- Meta-analysis shifts the focus from statistical significance to the *direction* and *magnitude* of the effect
- Key to this is the effect size
- It is the dependent variable of meta-analysis
- Encodes research findings on a numerical scale
- Different types of effect sizes for different research situations
- Each type may have multiple methods of computation

Overview

- Main types of effect sizes
- Logic of the standardized mean difference
- Methods of computing the standardized mean difference
- Logic of the odds ratio and risk ratio
- Methods of computing the odds ratio
- Adjustments, such as for baseline differences
- Issues related to the variance estimate

Most Common Effect Size Indexes

- Standardized mean difference (Cohen's *d* or Hedges' *g*)
  - Group contrast (e.g., treatment versus control)
  - Inherently continuous outcome construct
- Odds ratio and Risk ratio (*OR* and *RR*)
  - Group contrast (e.g., treatment versus control)
  - Inherently dichotomous (binary) outcome construct
- Correlation coefficient (*r*)
  - Two inherently continuous or scaled constructs

Less Common Effect Size Indexes

- Raw (unstandardized) mean difference
- Ratio of means
- Incident rate ratio for counts
- Proportion or prevalence (usually converted to a logit)
- Standardized gain score
- Standardized regression coefficient

Necessary Characteristics of Effect Sizes

- Numerical values produced must be comparable across studies
- Must be able to compute its standard error
- Must not be a direct function of sample size

The Standardized Mean Difference Effect Size

The Standardized Mean Difference

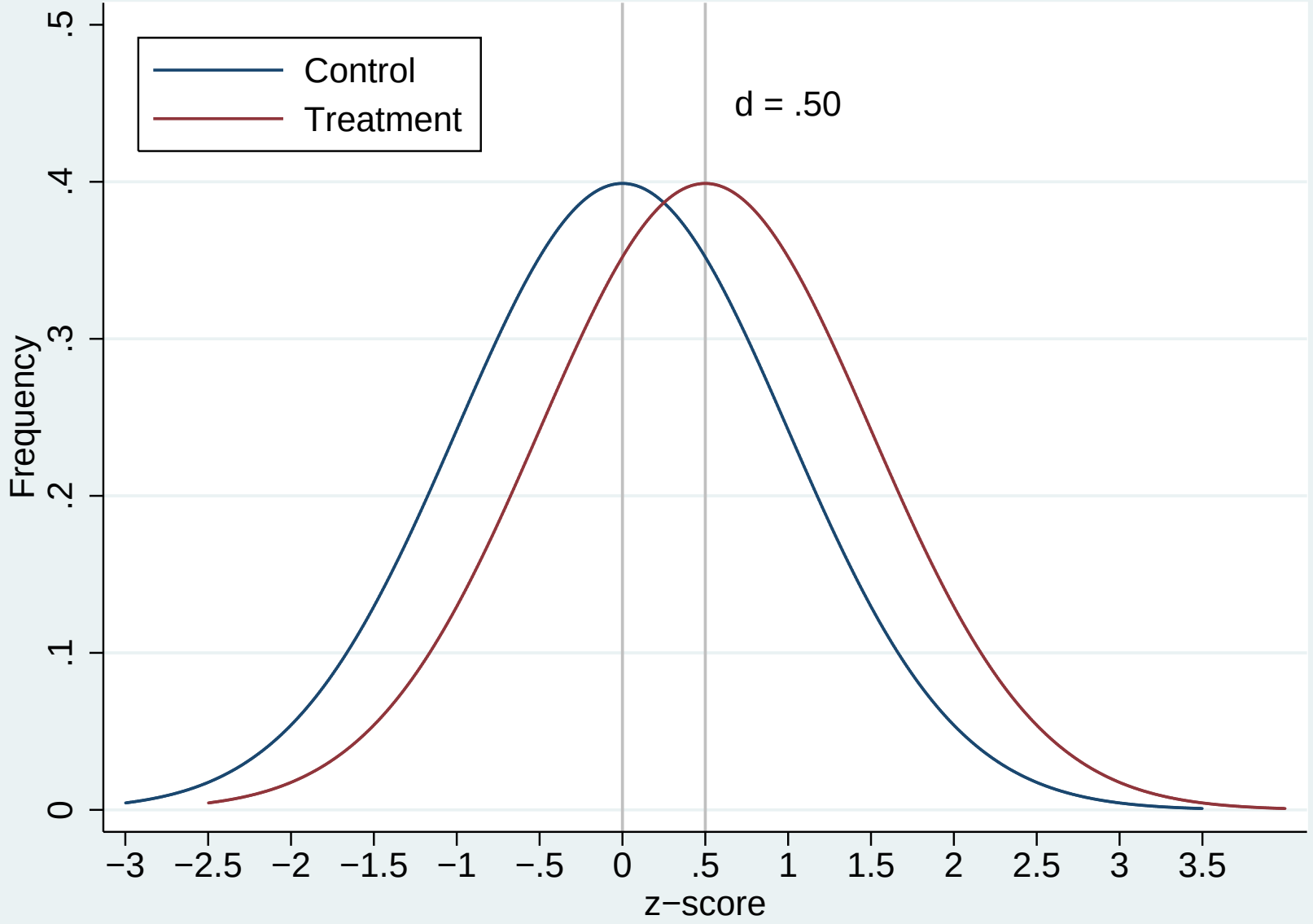
- Compares the means of two groups
- Standardizes this difference

$$d = \frac{\overline{X}_1 - \overline{X}_2}{s_{\text{pooled}}}$$

$$s_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}}$$

Notice that s_{pooled} is just a weighted average.

Visual Example of *d*-type Effect Size



Small Sample Size Bias Adjustment

- d is upwardly biased when based on small samples
- Bias is small when sample sizes in each condition are greater than 20
- Easy fix: Hedges' small sample size bias adjustment:

$$j = \left[1 - \frac{3}{4N - 9}\right]$$

$$j = \left[1 - \frac{3}{4(n_1 + n_2 - 2) - 1)}\right]$$

$$j = \left[1 - \frac{3}{4(df_1 + df_2) - 1)}\right]$$

- As the sample size increases, this adjustment approaches 1 (i.e., does not adjust).

Small Sample Size Bias Adjustment

To compute Hedges' g, we multiply Cohen's d by j

$$g = j \times d$$

It is common practice to always use this adjustment for standardized mean differences. However, if all of your studies have a large sample size (i.e., > 100), you can use Cohen's d.

Standard Error of d and g

The standard error (and variance) of d is mostly a function of sample size:

$$se_d = \sqrt{\left[\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2\,(n_1 + n_2)}\right]}$$

The standard error of g is the standard error of d with the small-sample-size bias adjustment applied. It also uses g instead of d in the equation.

$$se_g = j \times \sqrt{\left[\frac{n_1 + n_2}{n_1 n_2} + \frac{g^2}{2\,(n_1 + n_2)}\right]}$$

Variance of d and g

The variance of the effect size estimate is simply the squared standard error.

$$v_d = se_d^2$$

$$v_g = se_g^2$$

Methods of Computing *d*

- Lots of methods of computing *d*
- The goal is to reproduce what you would get with the means, standard deviations, and sample sizes
- Some methods are straightforward, others not
- Stata, R, and SPSS will compute *d* and *g*, assuming you have the mean, standard deviation, and sample size for each condition and for all studies, typically not the case.

Computing d from a t-test

Formula for d:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}$$

Formula for t:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

Therefore:

$$d = t\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

Computing *d* from a *p*-value from a *t*-test

- An exact *p*-value from a *t*-test can be convert into a *t*-value, assuming you have the degrees of freedom
- Then proceed as with the prior slide

Computing *d* from dichotomous outcomes

- Some studies may report a dichotomous outcome
- *d* (and *g*) can be estimated from these data
- Several methods
    1. Logit
    2. Cox Logit
    3. Probit
    4. Arcsine (no longer recommended)
- 1–3 have a similar logic and produce similar results unless the outcome has a low or high base rate (i.e., near 0 or 1).

Logit Method of Computing *d*

- Logistic distribution similar to the normal distribution
- log odds ratios conform to the logistic distribution
- Standard deviation of the logistic distribution is

$$s_{logistic} = \frac{\pi}{\sqrt{3}} = 1.814$$

- Thus, we can rescale a log *OR* into *d*

$$d = \frac{ln(OR)}{1.814}$$

- Monte Carlo simulations suggest that this performs fairly well
- However, there are some advantages to the Cox logit method (divide by 1.65)

$$d = \frac{ln(OR)}{1.65}$$

Probit Method

- The probit method uses the standard normal distribution rather than the logistic distribution
- Converts each proportion (i.e., successes in each group) into the corresponding *z*-score with *p* area under the left portion of the curve.
- Effect size is the difference between these two probits (z-values)

$$d = probit(p_1) - probit(p_2)$$

$$d = z_1 - z_2$$

Other Methods

There are lots of other methods for estimating *d* or *g*, depending on the data reported in the primary study. Most of these methods are built into my online calculator:

www.campbellcollaboration.org/calculator

Odds Ratio and Risk Ratio Effect Sizes

Logic of the Odds Ratio and Risk Ratio

- Odds ratio is the ratio of odds
- An odds is the probability of failure over the probability of success
- Risk ratio is the ratio of risks (probabilities)
- A risk is simply the probability of failure
- An odds ratio or risk ratio of 1 indicates no difference between the groups
- Odds ratios are more difficult to interpret than risk ratios
- Odds ratios are symmetrical to the outcome (this is why you can use them in case-control studies)

Logic of the Odds Ratio and Risk Ratio

Odds ratios and risk ratios contrast two groups on a dichotomous outcome

	Outcome	
	Success	Failure
Treatment	*a*	*b*
Control	*c*	*d*

where *a*, *b*, *c*, and *d*, are cell frequencies

Computing the Odds Ratio and Risk Ratio

The odds ratio is computed as:

$$OR = \frac{ad}{bc} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

The risk ratio is computed as:

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{p_1}{p_2}$$

Computing the Standard Error and Variance of the Odds ratio and Risk Ratio

Standard error of the log odds ratio:

$$se_{ln(OR)} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Standard error of the log risk ratio:

$$se_{\ln(RR)} = \sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}$$

Variance:

$$v = se^2$$

Computing the Odds Ratio from Other Data

- There are only so many ways you can report binary data
- Makes computing odds ratios and risk ratios relatively easy
- Generally have
  - Full two-by-two table
  - Percent of events (successes/failures) in each group
  - Proportion of events (successes/failures) in each group
  - Actual odds ratio or risk ratio
  - Convert *d*-type effect size

Convert *d*-type effect size to odds ratio

Just as we can convert an odds ratio in *d* to an odds ratio, we can convert reverse the process and convert *d* into an odds ratio:

- Using the logit method

$$\log OR = d\left(\frac{\pi}{\sqrt{3}}\right) = d \times 1.814 \tag{1}$$

- Using the Cox method

$$\log OR = d \times 1.65 \tag{2}$$

Prevalence Rate as Effect Sizes

Prevalence Rate as Effect Size

- Prevalence can be expressed as a proportion
- Logit of a proportion makes for a good effect size

$$\text{logit} = \ln\left[\frac{p}{1-p}\right]$$

- The standard error of the logit is

$$se_{\text{logit}} = \sqrt{\frac{1}{np} + \frac{1}{n(1-p)}}$$

- Final results can be back-converted into a proportion

$$p = \frac{e^{\text{logit}}}{e^{\text{logit}} + 1}$$

Using Stata, R, or SPSS to Calculate Effect Size

Using R, Stata, or SPSS to Calculate Effect Size

R, Stata, and SPSS can compute common effect size types from canonical data.

Using Stata to Calculate Effect Size

The syntax for Stata for the log odds ratio, log risk ratio, Hedges *g*, and Fisher's *Z* transformed *r* is:

```
* log odds ratio
meta esize a b c d, esize(lnoratio)

* log risk ratio
meta esize a b c d, esize(lnrratio)

* Hedges' g
meta esize n1 mean1 sd1 n2 mean2 sd2, esize(hedgesg)

* Convert r to Fisher's z
meta esize rho ssize, fisherz
```

Where a, b, c, and d are the variable names associated with the 2 by 2 frequency table. Similarly, n1, n2, mean1, mean2, sd1, and sd2 are the variable names for the sample sizes, means, and standard deviations for each group.

Using Stata to Calculate Effect Size

These commands add the following variables to your data set:

```
_meta_es
_meta_se
_meta_cil
_meta_ciu
_meta_studysize
```

Using R to Calculate Effect Size

For R, you have lots of effect size types you can compute. Below are code examples for the log odds ratio and log risk ratio:

```
library(metafor)
## log risk ratio from 2 by 2
escalc(measure="RR", ai=a, bi=b, ci=c, di=d, data=dat)

## log risk ratio from frequency positive and sample size
escalc(measure="RR", ai=a, n1i=n1, ci=c, n2i=n2, data=dat)

## log odds ratio from 2 by 2
escalc(measure="OR", ai=a, bi=b, ci=c, di=d, data=dat)

## log odds ratio from frequency positive and sample size
escalc(measure="OR", ai=a, n1i=n1, ci=c, n2i=n2, data=dat)
```

These functions return *yi* and *vi*, the effect size and its variance.

Using R to Calculate Effect Size

For R, you have lots of effect size types you can compute. Below is a code example for Hedges *g*.

```
## Hedges g
escalc(measure="SMD", m1i = mean1, m2i = mean2,
                      sd1i = sd1, sd2i = sd2,
                      n1i = n1, n2i = n2,
                      data = dat)
```

These functions return *yi* and *vi*, the effect size and its variance.

Using SPSS to Calculate Effect Size

For SPSS, the syntax for Hedges *g* is

```
* Hedges g
META CONTINUOUS
  /DATA TREATMENT = N(m1) MEAN(n1) STD(s1)
        CONTROL = N(m2) MEAN(n2) STD(s2)
        ID = ID STUDY=Authors
        ESTYPE=HEDGES_G(ADJUST)
  /SAVE ES(ES) SE_ES(seES) .
```

You can specify the names of the variables to add to the data set as shown in the SAVE line in the above examples.

Using SPSS to Calculate Effect Size

For SPSS, the syntax for the log odds ratio and log risk ratio is:

```
* log odds ratio
META BINARY
  /DATA TREATMENT=SUCCESS(a) FAILURE(b)
        CONTROL=SUCCESS(c) FAILURE(d)
        ID = ID STUDY=Authors
        ESTYPE=LOGOR
  /SAVE ES(ES) SE_ES(seES) .

* log risk ratio
META BINARY
  /DATA TREATMENT=SUCCESS(a) FAILURE(b)
        CONTROL=SUCCESS(c) FAILURE(d)
        ID = ID STUDY=Authors
        ESTYPE=LOGRR
  /SAVE ES(ES) SE_ES(seES) .
```

You can specify the names of the variables to add to the data set as shown in the SAVE line in the above examples.

Meta-analysis as a Systematic Review

Elements of a Systematic Review

- Explicit inclusion/exclusion criteria
- Exhaustive search for all eligible studies
- Systematic method of coding and data extraction
- Assessment of study quality
- Synthesis of the evidence (ideally using meta-analysis)

Eligibility criteria

Eligibility criteria need to be detailed.

- Treatment effectiveness meta-analyses often use PICO (population, intervention, comparator, and outcomes)
- In general, need to define:
  - Sample or population
  - Research design(s)
  - Measures
  - If treatment effectiveness: treatment and comparison conditions
  - Timeframe
  - Geographic/linguistic restrictions

Note: do not restrict to published, peer-reviewed studies.

Search for Studies

Search for studies should be exhaustive (and is often exhausting!)

- Use a trail-search coordinator or get assistance from a librarian, if possible
- Search multiple databases
- Search for grey literature
  - Dissertations/thesis
  - Conference presentations
  - Some databases include unpublished works
  - Hand search key journals
  - Contact experts in the field
  - Search Google Scholar

Screening Titles and Abstracts

- Titles and abstracts should be screened by two coders for potential eligibility
- Several software programs assist with this:
    1. Covidence
    2. Rayyan
    3. DistillerSR
    4. EPPI-Reviewer
    5. Abstrackr
    6. ASReview
- Most of these have built-in large-language models (LLMs) to facilitate screening (good area for AI use)
- Final eligibility should be based on a review of the full document, also done by two coders

Development of the Coding Protocol

- Like a survey but for studies
- A good coding protocol has transparency and replicability
- Goals:
  - Descriptive: characteristics of the studies, particularly ways in which the studies differ from each other
  - Extraction of findings: computation of effect sizes and related information

Assessment of Methodological Quality or Risk of Bias

- Cochrane has developed several widely used tools for assessing risk of bias.
    1. RoB 2: For person-based RCTs
    2. ROBINS-I: Designed to evaluate the risk of bias in non-randomized studies of interventions.
    3. ROBINS-E: Specifically tailored for non-randomized studies evaluating exposures.
    4. ROB-ME: Used to evaluate bias due to missing evidence (e.g., publication bias) in a synthesis
- No standard tools for the social sciences or ecology
- Learn what is commonly done in your field and tailor to your needs.

Data Structure

- Meta-analytic data is almost always hierarchical
  - Effect sizes nested within
  - Measures nested within
  - Samples or subsamples nested within
  - Studies
- Create separate coding forms and data tables for each level
- The idea is that of a relational database
- Must have identifying “keys” to connect the rows across tables
- Saves coding time, data cleanup time, and data manipulation time

Where to Code

- Paper forms
- Excel (common but has many weaknesses)
- Computer database (great choice!)
  - FileMaker
  - MS Access
  - REDCap (my new favorite)
- Specialized software
  - Comprehensive Meta-analysis
  - RevMan (Cochrane's tool)
  - Covidence
  - DistillerSR
  - EPPI-Reviewer

Double-Coding

- Reliability of coding is essential
- Best practice is to double-code everything and resolve any differences (consensus coding)
- What about AI as the second coder?

Basic Meta-Analysis

Overview of Method

- Goal:
  - Describe the distribution, including its mean
  - Establish a confidence interval around the mean
  - Test that the mean differs from zero
  - Test whether studies tell a consistent story (i.e., are homogeneous)
  - Explore the relationship between study features and effect size (i.e., explore heterogeneity)

Determining the Mean Effect Size

- Problem: studies have different sample sizes
- As such, effect sizes are more precise than other
- Maybe we should weight by sample size?
- What we need is an index of precision
- Actually, the standard error is a direct measure of precision
- Hedges and Olkin's solution:
  - Weight by the inverse variance (1 over the squared standard error)
- Provides a statistical basis for:
  - Standard error of the mean effect size
  - Confidence intervals
  - Homogeneity testing

Effect size transformations

- Not all effect sizes are analyzed in their raw form because the raw form does not have a standard error (or is problematic in some other way)
- For example:
  - Odds ratio –> log odds ratio
  - Risk ratio –> log risk ratio
  - Correlation –> Fisher's *r*-to-*z* transformation
  - Proportion –> logit
- Final results can be backtransformed into the original metric

Inverse Variance Weights

It is important to recognize the following equalities:

$$w = \frac{1}{v} = \frac{1}{se^2}$$

where

$$\begin{aligned} w &= \text{ inverse variance weight} \\ v &= \text{ variance} \\ se &= \text{ standard error} \end{aligned}$$

Most software programs (Stata, R, SPSS) will accept as input either the variance, standard error, or inverse variance weight.

Almost ready

- At this point, we have the following for each study:
  - An effect size
  - An inverse variance weight
- Problem: statistical models assume independence
- Only include one effect size per study (or independent sample)
- Multiple analyses for different subsets of independent effects
  - Different outcome constructs
  - Different time points
- We will discuss a method for analyzing dependent effect sizes (just not yet)

Inverse Variance Weighted Mean

Meta-analytic mean effect size is:

$$\overline{ES} = \frac{\sum w_i ES_i}{\sum w_i}$$

where ES_i is the effect size for each study (i) and w_i is the inverse variance weight

The standard error of the mean effect size is:

$$se_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}}$$

Some Basic Inferential Statistics

Confidence intervals can be constructed in the usual manner:

$$\overline{ES}_{lower} = \overline{ES} - se_{\overline{ES}}1.96$$
$$\overline{ES}_{upper} = \overline{ES} + se_{\overline{ES}}1.96$$

A z-test can be calculated as:

$$z = \frac{\overline{ES}}{se_{\overline{ES}}}$$

Example: Cognitive-Behavioral Programs for Adult Offenders

Author	Sample Size	*g* Effect Size	*w*
Burnett, 1996	60	0.45	5.684
Johnson & Hunter, 1995	98	0.11	15.934
Little & Robinson, 1989	180	0.23	25.049
Little et al 1991	152	0.22	27.852
Little et al 1994	1,381	0.34	150.485
Porporino & Robinson, 1995	757	0.04	65.529
Porporino et al 1991	63	0.16	11.953
Robinson, D., 1995	2,125	0.11	187.177
Ross et al 1988	45	1.28	6.441

Note: These studies are a subset of studies included in Wilson et al. (2005) and represent two specific treatment programs (Moral Reconation and Reasoning and Rehabilitation) and studies that were randomized or used high-quality quasi-experimental designs.

Computations by Hand

	g	w	$g \times w$
	0.45	5.684	2.55780
	0.11	15.934	1.75274
	0.23	25.049	5.76127
	0.22	27.852	6.12744
	0.34	150.485	51.16490
	0.04	65.529	2.62116
	0.16	11.953	1.91248
	0.11	187.177	20.58947
	1.28	6.441	8.24448
Sum	2.94	496.104	100.7317

$$\overline{g} = \frac{\sum gw}{\sum w} = \frac{100.7317}{496.104} = 0.203$$

$$se = \sqrt{\frac{1}{\sum w}} = \sqrt{\frac{1}{496.104}} = 0.045$$

$$z = \frac{\overline{g}}{se} = \frac{0.203}{0.045} = 4.511$$

$$\overline{g}_{lower95} = 0.203 - 1.96(0.045) = 0.115$$

$$\overline{g}_{upper95} = 0.203 + 1.96(0.045) = 0.291$$

Using Stata, R, or SPSS

Stata:

```
meta set g gse, fixed
meta summarize, nostudies
```

R using the *metafor* package:

```
rma(g, v, data=df, method="FE")
```

SPSS:

```
META ES CONTINUOUS
  /DATA ES = g VAR = v ID = studyid
  /INFERENCE MODEL = FIXED
  /PRINT HOMOGENEITY HETEROGENEITY.
```

Example: Cognitive-Behavioral Programs for Adult Offenders

Stata output

```
  Effect-size label: Effect size
        Effect size: d
          Std. err.: gse

Meta-analysis summary                                 Number of studies =      9
Fixed-effects model                                   Heterogeneity:
Method: Inverse-variance                                        I2 (%) =   43.62
                                                                H2 =    1.77

------------------------------------------------------------------------------
             |    Estimate    Std. err.     z    P>|z|    [95% conf. interval]
-------------+----------------------------------------------------------------
       theta |    .2045807    .0448966    4.56   0.000     .116585    .2925765
------------------------------------------------------------------------------
Test of homogeneity: Q = chi2(8) = 14.19                  Prob > Q = 0.0770
```

Example: Cognitive-Behavioral Programs for Adult Offenders

R output

```
Fixed-Effects Model (k = 9)

I^2 (total heterogeneity / total variability):   43.61%
H^2 (total variability / sampling variability):  1.77

Test for Heterogeneity:
Q(df = 8) = 14.1880, p-val = 0.0770

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.2030  0.0449  4.5225  <.0001  0.1150  0.2910  ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Cognitive-Behavioral Programs for Adult Offenders

Selected SPSS output

Effect Size Estimates

	Effect Size	Std. Error	Z	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
Overall	.203	.0449	4.522	<.001	.115	.291

Test of Homogeneity

	Chi-square (Q statistic)	df	Sig.
Overall	14.188	8	.077

Heterogeneity Measures

		Index	95% Confidence Interval	
			Lower	Upper
Overall	H-squared	1.773	.819	3.839
	I-squared (%)	43.6	.0	74.0

Homogeneity Testing

- Homogeneity analysis tests whether it is a reasonable assumption that all of the effect sizes are estimating the same population mean.
- If homogeneity is rejected, the distribution of effect sizes is assumed to be heterogeneous.
    - A single mean ES is not a good descriptor of the distribution
    - There are real differences between studies; that is, studies estimate different population means and effect sizes.
    - Three options:
        - model between study differences
        - fit a random effects model
        - do both

Homogeneity Q Statistic

- Q is simply a weighted sums-of-squares:

$$Q = \sum w_i (ES_i - \overline{ES})^2$$

- There are easier computational formulas:

$$Q = \sum w_i ES_i^2 - \frac{(\sum w_i ES_i)^2}{\sum w_i}$$

- It is distributed as a chi-square with $k - 1$ degrees of freedom, where k is the number of effect sizes

Higgins' I^2 Statistic

- Q is statistically underpowered when the number of studies is small, unless the sample sizes of those studies are large
- Jonathan Higgins proposed an alternative
- It is based on Q but doesn't focus on statistical significance

$$I^2 = \left(\frac{Q - df}{Q} \right) \times 100$$

- This reflects the percentage of the total variance in effects due to heterogeneity rather than sampling error
- Notice that this is a relative measure, not absolute. It is relative to the total variability across effect sizes.

Higgins' I^2 Statistic

Rough guide to interpreting I^2:

I^2	Excess variability
0% to 40%	might not be important
30% to 60%	may represent moderate heterogeneity
50% to 90%	may represent substantial heterogeneity
75% to 100%	may represent considerable heterogeneity

These are somewhat problematic as they presume that I^2 is an absolute, not relative, measure of heterogeneity.

H^2 Statistic

- An alternative to I^2
- H^2 is a ratio of total variability over sampling error variability (what would be expected if the distribution was homogeneous)
- It is an absolute measure of the amount of heterogeneity

$$H^2 = \frac{Q}{df}$$

- When H^2 equals 1, you have perfect homogeneity (observed variability across effect sizes equals the expected variability due to sampling error along)
- An H^2 of 2 would indicate twice the expected variability due solely to sampling error
- I'm not aware of any rules-of-thumb for interpreting H^2

Computations by Hand

	g	w	$g \times w$	$g^2 \times w$
	0.45	5.684	2.55780	1.1510
	0.11	15.934	1.75274	0.1928
	0.23	25.049	5.76127	1.3251
	0.22	27.852	6.12744	1.3480
	0.34	150.485	51.16490	17.3960
	0.04	65.529	2.62116	0.1048
	0.16	11.953	1.91248	0.3060
	0.11	187.177	20.58947	2.2648
	1.28	6.441	8.24448	10.5529
$\sum$	2.94	496.104	100.7317	34.6416

$$Q = \sum w_i ES_i^2 - \frac{(\sum w_i ES_i)^2}{\sum w_i}$$

$$Q = 34.6416 - -\frac{(100.7317)^2}{496.104}$$

$$Q = 34.6416 - 20.4531$$

$$Q = 14.19$$

$$df = 9 - 1 = 8, \quad p > .05$$

$$I^2 = 100 \times \frac{Q - df}{Q}$$

$$I^2 = 100 \times \frac{14.19 - 8}{14.19}$$

$$I^2 = 43.6\%$$

$$H^2 = 14.19/8 = 1.77$$

Example: Cognitive-Behavioral Programs for Adult Offenders

Stata output

```
  Effect-size label: Effect size
        Effect size: d
          Std. err.: gse

Meta-analysis summary                                 Number of studies =      9
Fixed-effects model                                   Heterogeneity:
Method: Inverse-variance                                        I2 (%) =   43.62
                                                                H2 =    1.77

------------------------------------------------------------------------------
             |    Estimate    Std. err.     z    P>|z|    [95% conf. interval]
-------------+----------------------------------------------------------------
       theta |    .2045807    .0448966    4.56   0.000     .116585    .2925765
------------------------------------------------------------------------------
Test of homogeneity: Q = chi2(8) = 14.19                  Prob > Q = 0.0770
```

Example: Cognitive-Behavioral Programs for Adult Offenders

R output

```
Fixed-Effects Model (k = 9)

I^2 (total heterogeneity / total variability):   43.61%
H^2 (total variability / sampling variability):  1.77

Test for Heterogeneity:
Q(df = 8) = 14.1880, p-val = 0.0770

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.2030  0.0449  4.5225  <.0001  0.1150  0.2910  ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Cognitive-Behavioral Programs for Adult Offenders

Selected SPSS output

Effect Size Estimates

	Effect Size	Std. Error	Z	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
Overall	.203	.0449	4.522	<.001	.115	.291

Test of Homogeneity

	Chi-square (Q statistic)	df	Sig.
Overall	14.188	8	.077

Heterogeneity Measures

		Index	95% Confidence Interval	
			Lower	Upper
Overall	H-squared	1.773	.819	3.839
	I-squared (%)	43.6	.0	74.0

Random Effects versus Fixed Effect Models

Random versus Fixed Effect Models

- The field is trying to change the label “fixed effect” to “common effect”
- Fixed effect model assumes:
  - there is one true population effect that all studies are estimating
  - all of the variability between effect sizes is due to sampling error
  - findings only apply to the set of studies included in the meta-analysis
- The random effects model assumes:
  - There are multiple (i.e., a distribution) population effects that the studies are estimating (i.e., the true effect size varies from study to study)
  - Variability between effect sizes is due to sampling error $+$ variability in the true population of effects

Fixed versus Random: Which to Use?

- A random-effects model becomes a fixed-effect model when distributions is homogeneous
- Assumptions of the fixed effects model are rarely plausible
  - Consequence: standard error that is too small; confidence intervals that are too narrow
- Bottom-line: Use the random-effects model (standard default) unless you can make a strong *a priori* case for a fixed effect model (i.e., strict replication studies)

Computing a Random Effects Model

- Fixed effects model: weights are a function of sampling error
- Random effects model: weights are a function of sampling error $+$ study level variability
- We call this study level variability τ^2
- Thus, we need a new set of weights

$$w = \frac{1}{se^2 + \tau^2}$$

Computing a Random Effects Model

- There are numerous methods of computing the random effects variance component, τ^2.
- Most of these require iterative methods and are not suitable for hand calculations.
- Below is the DerSimonian and Laird method-of-moments estimator, which can be easily computed by hand.
- First, compute τ^2 (random effects variance component):

$$\tau^2 = \frac{Q - df_Q}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}$$

- Second, recompute the inverse variance weights:

$$w_i = \frac{1}{se_i^2 + \tau^2}$$

- Third, recompute meta-analytic results using the new weight

Random Effects Computations by Hand

$$\tau^2 = \frac{Q - df_Q}{\sum w_i - \frac{\sum w_i^2}{\sum w_i}}$$

$$\tau^2 = \frac{14.19 - 8}{496.104 - \frac{63848.76}{496.104}}$$

$$\tau^2 = 0.0168$$

Random Effects Computations by Hand

Step 1. Compute v if you don't already have a column for it.

g	w	v
0.45	5.684	.175932
0.11	15.934	.062759
0.23	25.049	.039922
0.22	27.852	.035904
0.34	150.485	.006645
0.04	65.529	.015260
0.16	11.953	.083661
0.11	187.177	.005343
1.28	6.441	.155255

Random Effects Computations by Hand

Step 2. Add tau^2 to each v.

g	w	v	tau^2	$v + tau^2$
0.45	5.684	.175932	.0168	.192732
0.11	15.934	.062759	.0168	.079559
0.23	25.049	.039922	.0168	.056722
0.22	27.852	.035904	.0168	.052704
0.34	150.485	.006645	.0168	.023445
0.04	65.529	.015260	.0168	.032060
0.16	11.953	.083661	.0168	.100461
0.11	187.177	.005343	.0168	.022143
1.28	6.441	.155255	.0168	.172055

Random Effects Computations by Hand

Step 3. Compute new weight $(1/(v + tau^2)$. In this example, these are DerSiminian and Laird weights (hence the "DL")

g	w	v	tau^2	$v + tau^2$	w_{DL}
0.45	5.684	.175932	.0168	.192732	5.18854
0.11	15.934	.062759	.0168	.079559	12.56931
0.23	25.049	.039922	.0168	.056722	17.62992
0.22	27.852	.035904	.0168	.052704	18.97387
0.34	150.485	.006645	.0168	.023445	42.65269
0.04	65.529	.015260	.0168	.032060	31.19111
0.16	11.953	.083661	.0168	.100461	9.95411
0.11	187.177	.005343	.0168	.022143	45.16195
1.28	6.441	.155255	.0168	.172055	5.81208

Random Effects Computations by Hand

Steps 4 and 5. Multiply the new weights by the effect size and sum the columns.

	g	w	v	tau^2	$v + tau^2$	w_{DL}	$g \times w_{DL}$
	0.45	5.684	.175932	.0168	.192732	5.18854	2.3348
	0.11	15.934	.062759	.0168	.079559	12.56931	1.3826
	0.23	25.049	.039922	.0168	.056722	17.62992	4.0549
	0.22	27.852	.035904	.0168	.052704	18.97387	4.1743
	0.34	150.485	.006645	.0168	.023445	42.65269	14.5019
	0.04	65.529	.015260	.0168	.032060	31.19111	1.2476
	0.16	11.953	.083661	.0168	.100461	9.95411	1.5927
	0.11	187.177	.005343	.0168	.022143	45.16195	4.9678
	1.28	6.441	.155255	.0168	.172055	5.81208	7.4395
Sum	2.94					189.1336	41.6961

Random Effects Computations by Hand

Step 6. Recompute mean effect size and standard error as before.

Comment:

Because the variance is larger, the weight is smaller. Also, across studies, the weights will be more similar. As τ^2 approaches infinity, the weights will become virtually equal (i.e., τ^2 never gets this big!).

Using Stata, R, or SPSS

Stata:

```
meta set g gse, random(dl)
meta summarize
```

R using the *metafor* package:

```
rma(g, v, data=df, method="DL")
```

SPSS:

```
META ES CONTINUOUS
  /DATA ES = g VAR = v ID = studyid
  /INFERENCE MODEL = RANDOM ESTIMATE = DERSIMONIAN_LAIRD
  /PRINT HOMOGENEITY HETEROGENEITY.
```

Example: Cognitive-Behavioral Programs for Adult Offenders

R output:

```
Random-Effects Model (k = 9; tau^2 estimator: DL)

tau^2 (estimated amount of total heterogeneity): 0.0168 (SE = 0.0216)
tau (square root of estimated tau^2 value):      0.1298
I^2 (total heterogeneity / total variability):   43.61%
H^2 (total variability / sampling variability):  1.77

Test for Heterogeneity:
Q(df = 8) = 14.1880, p-val = 0.0770

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.2205  0.0728  3.0304  0.0024  0.0779  0.3631  **

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estimators for τ^2

There are numerous estimators for the random effects variance component.

- Dersimonian-Laird method-of-moments is (has been) the most widely used (but shouldn't be)
- Doesn't always perform well when the number of studies is small and heterogeneity is high
- Restricted maximum likelihood (REML)
  - Widely recommended alternative
  - Requires iteration; not suitable for hand calculation
  - Is now widely available in most software programs (and is the default in many)
- Metafor package in R implements 12 estimators
- Stata estimates 7
- SPSS also estimates 7

Comparison of Random Effects and Fixed Effect Models

- The biggest difference you will notice is the significance levels and confidence intervals
- As heterogeneity increases
  - The *p*-value for the mean effect size will increase
  - The confidence interval will get wider
- A mean that was significant under a fixed effect model might not be significant under the random effects model
- The random effects model is more conservative
- Mean effect size may differ between models if sample sizes and effect sizes are related

Summary of Basic Meta-analysis

- Apply any needed transformations to effect sizes
  - Fisher's *Zr* transformation (*r*)
  - Natural log (*OR* and *RR*).
- Compute inverse variance weight
- Compute mean, standard error, and confidence intervals
- Compute homogeneity statistics (Q, I^2, H^2)
- Assume random effects unless you can make a strong *a priori* case for a fixed-effect model

Prediction Intervals

Prediction Intervals

Beyond Confidence Intervals: Prediction Intervals

Confidence interval:

- Estimates uncertainty about the *mean* effect
- "Where is the average effect likely to be?"
- Narrows as more studies are added

Prediction interval:

- Estimates range for a *new study's* effect
- "Where might the next study's effect fall?"
- Accounts for both sampling error *and* heterogeneity (τ^2)
- Does not narrow much as studies are added (if heterogeneity exists)

Why it matters:

- More realistic representation of uncertainty
- Critical when heterogeneity is substantial
- Helps contextualize clinical/practical significance

Computing Prediction Intervals

Formula for 95% prediction interval:

$$PI = \overline{ES} \pm 1.96 \times \sqrt{se^2_{\overline{ES}} + \tau^2}$$

where:

- $\overline{ES}$ = mean effect size
- 1.96 critical z value, 95% region
- $se_{\overline{ES}}$ = standard error of mean effect
- τ^2 = between-study variance

Key difference from CI:

- CI uses: $\sqrt{se^2_{\overline{ES}}}$
- PI uses: $\sqrt{se^2_{\overline{ES}} + \tau^2}$
- The $+\tau^2$ makes PI wider when heterogeneity exists

Interpreting Prediction Intervals

Example: Treatment effect on depression

- Mean effect: $g = 0.50$ (95% CI: 0.35 to 0.65)
- Prediction interval: 95% PI: -0.10 to 1.10

Interpretation:

- We're confident the *average* effect is positive (CI excludes zero)
- But a new study might find effects ranging from slightly negative to very large
- This reflects real heterogeneity across contexts
- Important for generalization and implementation

When PI includes zero:

- Even if the mean is significant, some contexts may show no effect
- Suggests the need for moderator analysis
- Caution about universal effectiveness claims

Prediction Intervals in R

Using metafor:

```
# Fit random-effects model
mod <- rma(g, v, data = cbt, method = "REML")
predict(mod, digits = 3)
```

Output includes:

```
   pred     se  ci.lb  ci.ub   pi.lb  pi.ub
 0.2139 0.0634 0.0897 0.3382 -0.0133 0.4411
```

Prediction Intervals in Stata

Stata:

```
meta summarize, predinterval
```

```
 Effect-size label: Effect size
       Effect size: g
         Std. err.: gse

Meta-analysis summary                                 Number of studies =      9
Random-effects model                                  Heterogeneity:
Method: REML                                                     tau2 =  0.0094
                                                        I2 (%) =   30.20
                                                            H2 =    1.43

------------------------------------------------------------------------------
             |    Estimate    Std. err.     z    P>|z|    [95% conf. interval]
-------------+----------------------------------------------------------------
       theta |    .2139363     .063385    3.38   0.001     .089704    .3381685
------------------------------------------------------------------------------
95% prediction interval for theta: [-0.060, 0.488]

Test of homogeneity: Q = chi2(8) = 13.85                  Prob > Q = 0.0859
```

Prediction Intervals in SPSS

In SPSS, add `PREDICTION` to the `PRINT` subcommand.

```
META ES CONTINUOUS
  /DATA ES = g VAR = v ID = studyid
  /INFERENCE MODEL = RANDOM ESTIMATE = DERSIMONIAN_LAIRD
  /PRINT HOMOGENEITY HETEROGENEITY PREDICTION.
```

This adds the 95% prediction interval to the Effect Size Estimates table.

Prediction Intervals: Key Takeaways

1. **PI $\neq$ CI**: They answer different questions
    - CI: Where is the average effect?
    - PI: Where might a new study's effect be?
2. **PI width reflects heterogeneity**
    - Narrow PI: consistent effects across studies
    - Wide PI: substantial variation in effects
3. **Clinical/practical significance**
    - PI crossing zero suggests variable effectiveness
    - Motivates moderator analysis
4. **Increasingly expected by journals**
    - Cochrane Handbook recommends reporting PI
    - More transparent about uncertainty

Forest Plots

Forest Plots

- Visual representation of results
- Row for each study that shows
  - study label
  - sample size; may include other information
  - effect size (dot, square, diamond)
  - confidence interval (horizontal line)
- Row for the overall mean results
  - effect size (dot, square, diamond)
  - confidence interval (horizontal line)

Forest Plots in Stata, R, and SPSS

Stata:

```
meta set g gse, random(reml) studylabel(author)
meta forest
```

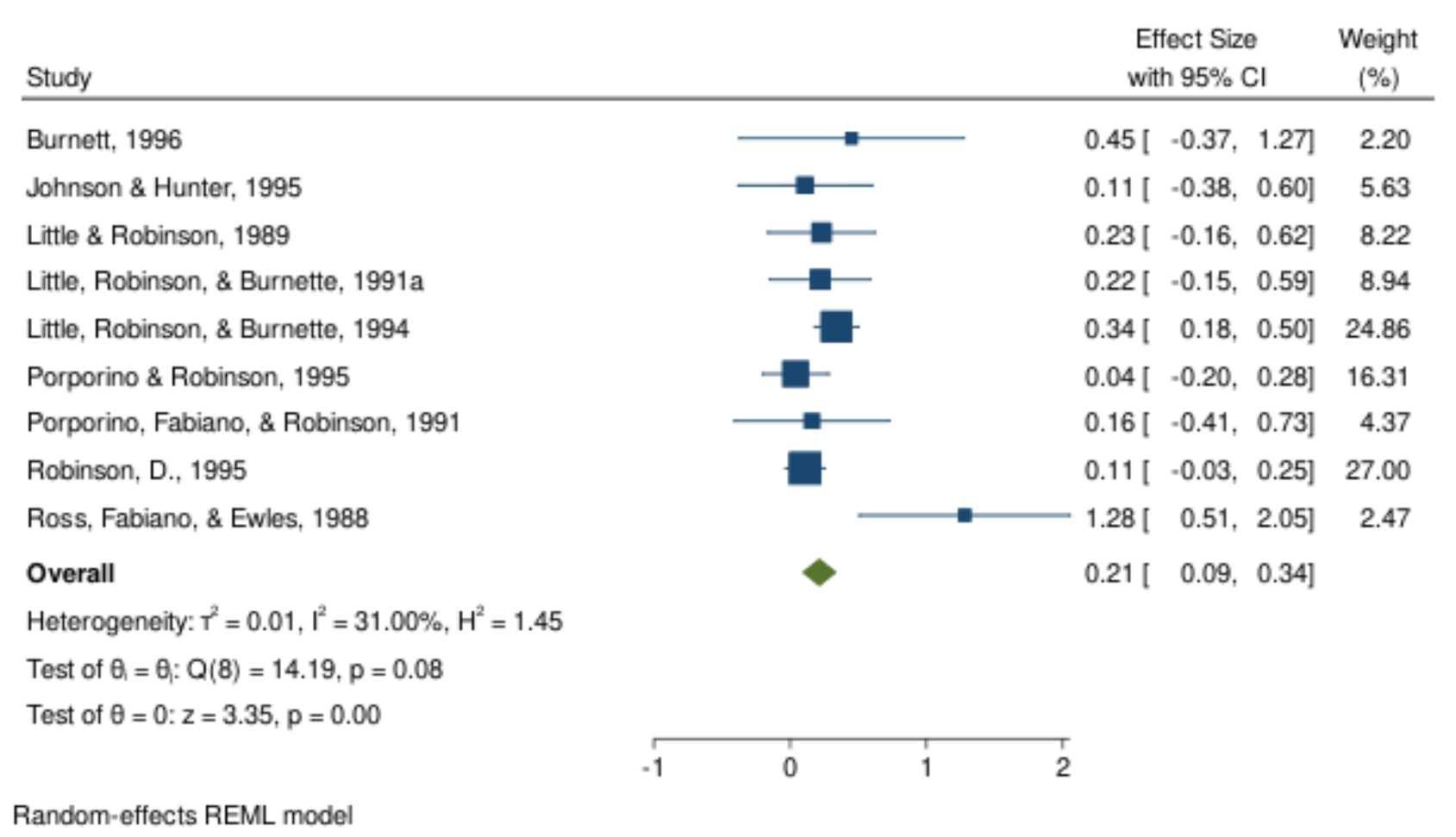
R using the *metafor* package:

```
mod <- rma(g, v, data=df,
           method="REML", slab=author)
forest(mod)
```

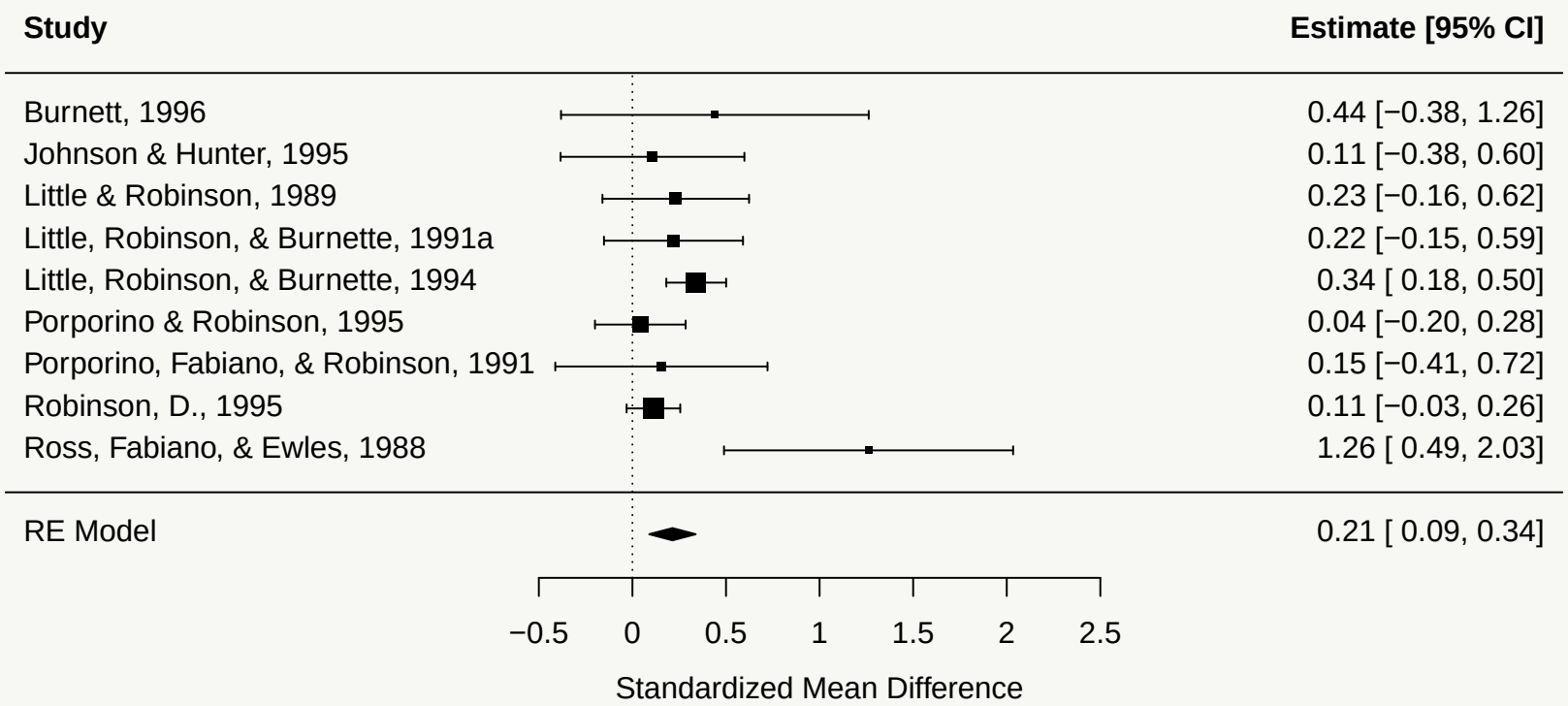
SPSS:

```
META ES CONTINUOUS
   /DATA ES = g VAR = v STUDY = author
   /INFERENCE MODEL = RANDOM ESTIMATE = REML
   /FORESTPLOT DISPLAY = ES CI WEIGHT
               POSITION = RIGHT REFLINES = NULL.
```

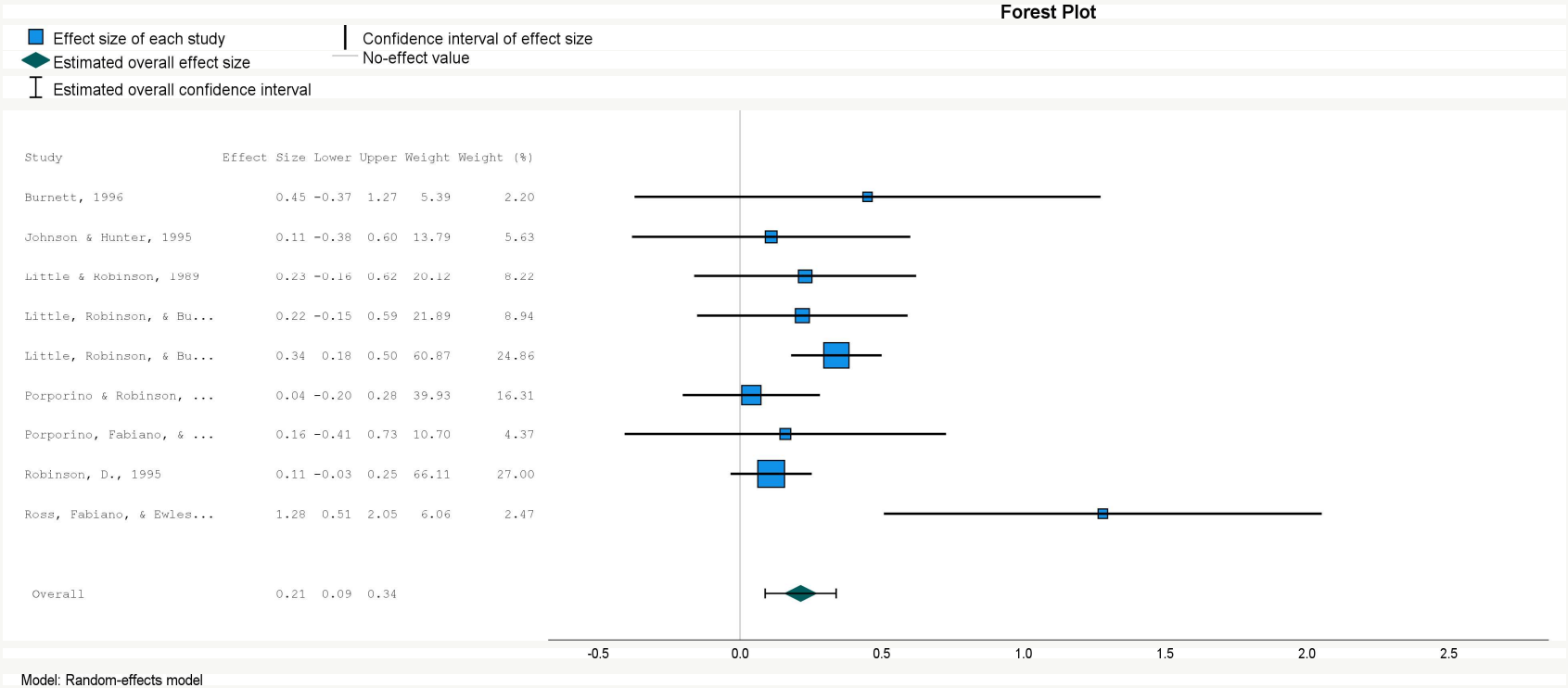
Forest Plot Stata Example



Forest Plot R Example



Forest Plot SPSS Example



Moderator Analysis

Moderator Analysis

- Modeling between study variability
  - Categorical models (subgroup analysis; analogous to a one-way ANOVA)
  - Regression models
- Fixed and random effects versions of each (latter often called “mixed” models)
- Number of studies needed

Subgroup Analysis: Analog to the ANOVA

Three (related) Approaches

There are different approaches to conducting a moderator analysis on a categorical variable:

1. Run separate basic meta-analyses on subgroups of studies
    - Pro: Easy to do and understand
    - Con: Doesn't provide a test for moderation (i.e., do the means differ across categories)
2. Perform an analog to the ANOVA type of analysis
3. Perform a meta-regression using dummy codes

Depending on the method, you will either estimate separate τ^2s for each category or a common τ^2 across categories.

Analog to the ANOVA

- Useful for a single categorical independent variable
- Produce a separate mean effect size for each category
- Also produces a test of whether the mean effect sizes differ across categories.
- Recall that Q is a sum-of-squares
- The total sum-of-squares (Q) can be partitioned
  - Variability between groups (Q_{Between})
  - Residual variability within groups (Q_{Within})

Analog to the ANOVA

- Q_{Between} (might also be called Q_{Model}) analogous to an F-test between means
- Q_{Within} (might also be called Q_{Residual}) assesses whether residual distribution homogeneous

Analog to the ANOVA Equations

Recall that the overall (or total) Q is computed as

$$Q_{Total} = \sum w_i (ES_i - \overline{ES})^2$$

$$Q_{Total} = \sum w_i {ES_i}^2 - \frac{\left(\sum w_i ES_i\right)^2}{\sum w_i}$$

Analog to the ANOVA Equations

To compute the Q_{Within}, we need to compute the Q within each group and then sum across groups

$$Q_{\text{Within}} = \sum w_{ij}(ES_{ij} - \overline{ES_j})^2$$

Analog to the ANOVA Equations

To compute the Q_{Between}, simply subtract Q_{Within} from Q_{Total}

$$Q_{\text{Between}} = Q_{Total} - Q_{\text{Within}}$$

Recall that Q is distributed as a χ^2. The degrees of freedom are computed just as they are in a one-way ANOVA:

$$df_{\text{Within}} = k - j$$

$$df_{\text{Between}} = j - 1$$

$$df_{Total} = k - 1$$

where k is the number of effect sizes and j is the number of categories or groups.

Analog to the ANOVA Interpretation

- Q_{Between} reflects the differences in the means across categories (analogous to the F from a one-way ANOVA
- Q_{Within} indicates whether the distributions within categories are homogeneous (overall)
    - For the random effects model, Q_{Within} should be based on a fixed effect model or a significance test of τ^2
- The test is often statistically under-powered (particularly in the random effects case)
- The random effects version estimates τ^2 based on the within groups heterogeneity (i.e., it will be smaller than for the full model)

Using Stata, R, or SPSS

Stata:

```
meta set g gse, random(reml)
meta summarize, subgroup(random)
```

R using the *metafor* package:

```
## run dummy coded regression model to get Q_between (QM)
model <- rma(g ~ factor(random), v, data=df, method="REML")
print(model)
## run the same model without intercept to get subgroup means
model <- rma(g ~ 0 + factor(random), v, data=df, method="REML")
print(model)
```

SPSS:

```
META ES CONTINUOUS
   /DATA ES = g VAR = v ID = studyid
   /ANALYSIS SUBGROUP = random
   /INFERENCE MODEL = RANDOM ESTIMATE = REML
   /PRINT HOMOGENEITY HETEROGENEITY.
```

In these examples, “random” is our moderator variable.

Analog to the ANOVA in Stata (part 1)

```
  Effect-size label: Effect size
        Effect size: g
          Std. err.: gse

Subgroup meta-analysis summary                     Number of studies =      9
Random-effects model
Method: REML
Group: random

------------------------------------------------------------------------------
             |       theta    Std. err.     z    P>|z|    [95% conf. interval]
-------------+----------------------------------------------------------------
          no |    .2305873    .1190659    1.94   0.053   -.0027775    .4639521
         yes |    .2266274    .1017047    2.23   0.026    .0272897     .425965
-------------+----------------------------------------------------------------
     Overall |    .2139363     .063385    3.38   0.001     .089704    .3381685
------------------------------------------------------------------------------
```

Analog to the ANOVA in Stata (part 2)

```
Heterogeneity summary
------------------------------------------------------------------------------
        Group |      df          Q   P > Q        tau2       % I2         H2
--------------+---------------------------------------------------------------
           no |       3       0.32   0.956    1.61e-08       0.00       1.00
          yes |       4      13.46   0.009    .0287469      66.74       3.01
--------------+---------------------------------------------------------------
      Overall |       8      13.85   0.086    .0094198      30.20       1.43
------------------------------------------------------------------------------
Test of group differences: Q_b = chi2(1) = 0.00           Prob > Q_b = 0.980
```

Software Differences

- Stata and SPSS estimate separate τ^2 using the methods shown above
- R using the `metafor` package is estimating a common τ^2. For separate *tau*^{2} values, run separate models (or use the `meta` package.

Meta-analytic Regression

Meta-analytic Regression

- Conceptually the same as multiple regression
  - Effect size is the dependent variable
  - Study moderator variables are the independent variables
- Can handle multiple variables simultaneously
- Don’t use standard OLS regression procedures
- UWLS (unrestricted weighting by inverse variance) is used in economics (see work by TD Stanley)
- Must use specialized software

Using Stata, R, or SPSS

Stata:

```
meta set g gse, random(reml)
meta regress random
```

R using the *metafor* package:

```
model <- rma(g ~ random, v, data=df, method="REML")
print(model)
```

SPSS:

```
META REGRESSION g WITH random
   /DATA VAR = v
   /INFERENCE MODEL = RANDOM ESTIMATE = REML
              DISTRIBUTION = NORMAL
   /PRINT PARAMETER.
```

Meta-analytic Regression (Stata)

```
      Effect-size label:  Effect Size
        Effect size:  d
          Std. Err.:  gse

Random-effects meta-regression                  Number of obs  =         9
Method: REML                                    Residual heterogeneity:
                                                           tau2 =   .01286
                                                         I2 (%) =    37.56
                                                            H2 =     1.60
                                                R-squared (%) =     0.00
                                                Wald chi2(1)  =     0.02
                                                Prob > chi2   =   0.8888
------------------------------------------------------------------------------
    _meta_es |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      random |  -.0218312    .156078    -0.14   0.889    -.3277384     .284076
       _cons |   .2345103   .1347075     1.74   0.082    -.0295114    .4985321
------------------------------------------------------------------------------
Test of residual homogeneity: Q_res = chi2(7) = 14.13    Prob > Q_res = 0.0490
```

Meta-analytic Regression (R)

```
Mixed-Effects Model (k = 9; tau^2 estimator: REML)

tau^2 (estimated amount of residual heterogeneity):     0.0129 (SE = 0.0194)
tau (square root of estimated tau^2 value):             0.1134
I^2 (residual heterogeneity / unaccounted variability): 37.56%
H^2 (unaccounted variability / sampling variability):   1.60
R^2 (amount of heterogeneity accounted for):            0.00%

Test for Residual Heterogeneity:
QE(df = 7) = 14.1264, p-val = 0.0490

Test of Moderators (coefficient 2):
QM(df = 1) = 0.0196, p-val = 0.8888

Model Results:

         estimate      se     zval    pval    ci.lb   ci.ub
intrcpt    0.2345  0.1347   1.7409  0.0817  -0.0295  0.4985
random    -0.0218  0.1561  -0.1399  0.8888  -0.3277  0.2841
```

Meta-analytic Regression (SPSS)

Selected SPSS output

Parameter Estimates

Parameter	Estimate	Std. Error	Z	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
(Intercept)	.234	.1348	1.738	.082	-.030	.499
use of random assignment to conditions	-.023	.1563	-.148	.882	-.329	.283

Test of Residual Homogeneity

Chi-square (Q statistic)	df	Sig.
14.119	7	.049

Tests the null hypothesis that tau-squared is equal to 0.

Residual Heterogeneity

Tau-squared	.013
I-squared (%)	37.7
H-squared	1.606
R-squared (%)	.0

Dependent Effect Sizes

The Problem of Dependent Effect Sizes

Traditional meta-analysis assumes independence

- One effect size per study
- But real studies often report:
  - Multiple outcomes (depression, anxiety, quality of life)
  - Multiple time points (post-test, 3-month, 6-month follow-up)
  - Multiple treatment arms (3 treatments vs. control)
  - Multiple subgroups (boys vs. girls, young vs. old)

The traditional "solution": Select one effect size per study

- Wastes data
- Arbitrary decisions reduce replicability
- Loses statistical power

Modern Solutions for Dependent Effect Sizes

Three main approaches:

1. **Robust Variance Estimation (RVE)**
   - Accounts for unknown correlation structure
   - Cluster-robust standard errors
   - Requires fewer assumptions
2. **Three-level (multilevel) meta-analysis**
   - Models the dependency structure explicitly
   - Level 1: sampling variance
   - Level 2: within-study variance
   - Level 3: between-study variance
   - Makes strong assumptions (i.e., simple dependence structure)
3. **Multivariate meta-analysis**
   - Requires known correlations between effect sizes
   - Most appropriate for different outcomes

Robust Variance Estimation (RVE): The Basics

Key idea: Adjust standard errors to account for clustering

- Developed by Hedges, Tipton, & Johnson (2010)
- Uses a "working" correlation model
- Produces cluster-robust standard errors
- Common correlation assumption: $\rho = 0.8$ (sensitivity analysis recommended)

Small sample correction:

- Original RVE: needs $k \geqslant 40$ studies
- Small-sample corrections available (Tipton, 2015; Tipton & Pustejovsky, 2015)
- Use degrees of freedom adjustment (similar to Satterthwaite)

RVE in R: robu() Function

R code using robumeta package:

```
library(robumeta)
# Basic RVE model
rve_model <- robu(effectsize ~ 1,
                  data = dat,
                  studynum = studyid,
                  var.eff.size = variance,
                  rho = 0.8,
                  small = TRUE)  # small-sample correction
summary(rve_model)

# RVE with moderators
rve_mod <- robu(effectsize ~ moderator1 + moderator2,
                data = dat,
                studynum = studyid,
                var.eff.size = variance,
                rho = 0.8,
                small = TRUE)
summary(rve_mod)
```

RVE in Stata

Stata code using robumeta:

```
* Installs robumeta (only need to run once)
ssc install robumeta

* Basic RVE model
robumeta es, var.eff.size(variance) studynum(studyid) rho(0.8)

* RVE with moderators
robumeta es moderator1 moderator2, ///
      var.eff.size(variance) studynum(studyid) rho(0.8)
```

Publication Selection Bias

Publication Selection Bias

- Statistically significant effects are more likely to be published than nonsignificant effects
- Has many forms
  - outcomes selected for inclusion in a report
  - analyses selected for inclusion in a report
  - author's desire to write up a study
  - reviewers' recommendations for publication
  - editor's decision to accept for publication
- Existence of this bias is well established
- Affects all forms of reviewing the literature
- Essential that you search for and include unpublished studies that meet eligibility criteria
- It is essential that you explore the risk of publication selection bias in your data

Evolution of Publication Bias Methods

Traditional methods (still useful but limited):

- Funnel plots (visual inspection)
- Egger's regression test
- Trim-and-fill
- Fail-safe N (not recommended)

Evolution of Publication Bias Methods

What reviewers are now asking for:

- Selection models
- PET-PEESE (Precision-Effect Test/Estimate with Standard Error)
- P-curve or p-uniform
- Multiple methods as sensitivity analyses

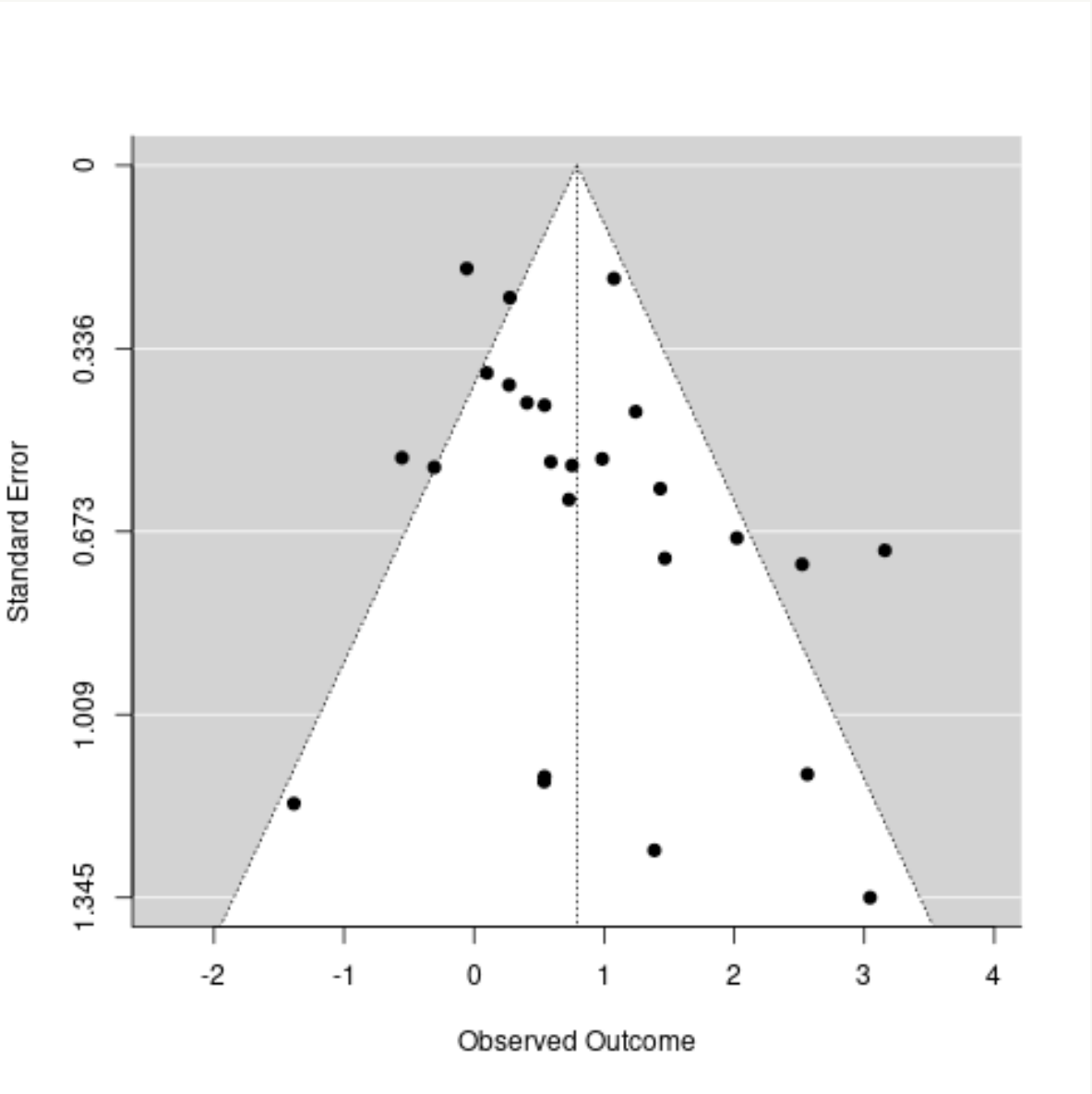
Why the shift?

- Trim-and-fill performs poorly in many scenarios
- Egger's test has low power and can be influenced by heterogeneity
- Newer methods have better statistical properties

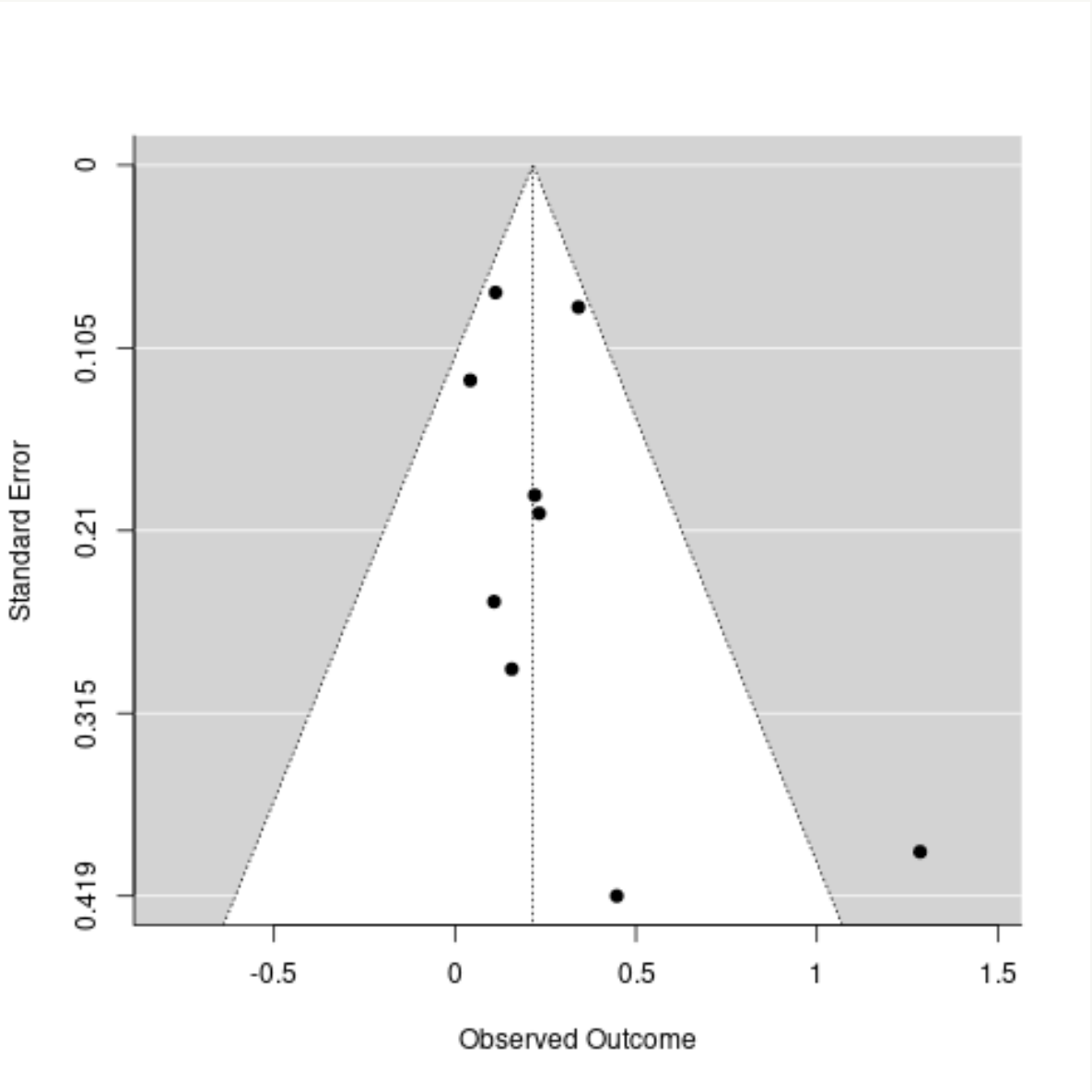
Funnel Logic

- As sample size increases, smaller effects become statistically significant
- Thus, smaller effects from smaller studies are more likely to be missing
- Statistical significance is directly a function of the standard error
- Funnel plot is a scatter plot if the inverse of the standard error (large values reflect larger samples) and effect size
- The expectation is that the plot should be symmetric around the mean effect size
- I.e., it should look like an upside-down funnel
- Asymmetry is suggestive of publication selection bias

Funnel Plot Example 1



Funnel Plot Example 2



Funnel Plot in Stata, R, and SPSS

Stata:

```
meta set g gse, random(reml)
meta funnel
```

R using the *metafor* package:

```
mod <- rma(g, v, data=df, method="REML")
funnel(mod)
```

SPSS:

```
META ES CONTINUOUS
    /DATA ES = g VAR = v STUDY = AUTHOR author
    /INFERENCE MODEL = RANDOM ESTIMATE = REML
    /FUNNELPLOT YAXIS=SE.
```

Trim and Fill Method

- Method for assessing the impact of publication selection bias on results
- Based on the same idea as the funnel plot
- Nonparametric method for adjusting the mean effect size
- Rectifies the funnel plot asymmetry
  - Assumes the effects to the left of the mean are more likely to be missing
  - Initially trims effect sizes iteratively to make the funnel plot symmetric
  - Replaces trimmed effect sizes with pairs on the left side of the mean

Trim and Fill in Stata, R, and SPSS

Stata:

```
meta set g gse, random(reml)
meta trimfill, left
```

R using the *metafor* package:

```
mod <- rma(g, v, data=df, method="REML")
trimfill(mod, "left")
```

SPSS:

```
META ES CONTINUOUS
   /DATA ES = g VAR = v STUDY = author
   /INFERENCE MODEL = RANDOM ESTIMATE = REML
   /TRIMFILL SIDE=LEFT.
```

Trim and Fill Method, Stata output

```
Nonparametric trim-and-fill analysis of publication bias
Linear estimator, imputing on the left

Iteration                              Number of studies =     11
  Model: Random-effects                        observed =      9
 Method: REML                                  imputed =      2

Pooling
  Model: Random-effects
 Method: REML

------------------------------------------------------------------
             Studies |      Effect Size    [95% Conf. Interval]
---------------------+--------------------------------------------
            Observed |             0.215       0.090       0.340
  Observed + Imputed |             0.181       0.048       0.315
------------------------------------------------------------------
```

Trim and Fill Method, R output

```
Estimated number of missing studies on the left side: 2 (SE = 2.1217)

Random-Effects Model (k = 11; tau^2 estimator: REML)

tau^2 (estimated amount of total heterogeneity): 0.0144 (SE = 0.0191)
tau (square root of estimated tau^2 value):      0.1198
I^2 (total heterogeneity / total variability):   35.46%
H^2 (total variability / sampling variability):  1.55

Test for Heterogeneity:
Q(df = 10) = 22.6187, p-val = 0.0122

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.1812  0.0681  2.6626  0.0078  0.0478  0.3146  **
```

Trim and Fill Method, SPSS output

Effect Size Estimates for Trim-and-Fill Analysis

	Number	Effect Size	Std. Error	Z	Sig. (2-tailed)	95% ... Lower
Observed	9	.214	.0639	3.350	<.001	.089
Observed + Imputed[a]	11	.180	.0681	2.642	.008	.046

Effect Size Estimates for Trim-and-Fill Analysis

	95% ... Upper
Observed	.339
Observed + Imputed[a]	.313

a. Number of imputed studies: 2

Egger's Test for Funnel-Plot Asymmetry

- Egger's test provides a method of statistically assessing for funnel plot asymmetry
- It regresses the effect size on its precision, the standard error. This model is also weighted by the inverse of the variance.
- The regression coefficient has no clear interpretation
- Interpretation focuses on statistical significance
- Stata and R agree on the significance test
- Stata reports the slope, R reports the significance of the slope, and SPSS reports both (but it doesn't tell you which one to interpret!)
- Can be done under both a fixed-effect and random-effects model (you will get different results)
- There are reasons other than publication bias that may cause asymmetry

Egger's Test in Stata, R, and SPSS

Stata:

```
meta set g gse, random(reml)\
meta bias, egger
```

R using the *metafor* package:

```
mod <- rma(g, v, data=df, method="REML")
regtest(mod)
```

SPSS:

```
META ES CONTINUOUS
   /DATA ES = g VAR = v STUDY = author
   /INFERENCE MODEL = RANDOM ESTIMATE = REML
   /BIAS INTERCEPT=INCLUDE DISTRIBUTION = NORMAL.
```

Egger's Test, Stata output

```
Regression-based Egger test for small-study effects
Random-effects model
Method: REML

H0: beta1 = 0; no small-study effects
            beta1 =      1.08
      SE of beta1 =     0.769
                z =      1.40
       Prob > |z| =    0.1619
```

Egger's Test, R output

```
Regression Test for Funnel Plot Asymmetry

Model:     mixed-effects meta-regression model
Predictor: standard error

Test for Funnel Plot Asymmetry: z = 1.3987, p = 0.1619
Limit Estimate (as sei -> 0):   b = 0.0625 (CI: -0.1962, 0.3211)
```

Egger's Test, SPSS output

Egger's Regression-Based Test[a]

Parameter	Coefficient	Std. Error	Z	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
(Intercept)	.060	.1321	.453	.650	-.199	.319
SE[b]	1.084	.7693	1.410	.159	-.423	2.592

a. Random-effects meta-regression

b. Standard error of effect size

PET-PEESE: Precision-Effect Test and Estimate

PET (Precision-Effect Test):

- Regress effect size on standard error (same as Egger's test, but focuses on the intercept, not the slope)
- If intercept = 0, no evidence of genuine effect beyond bias
- Used as a test of publication bias

PEESE (Precision-Effect Estimate with Standard Error):

- Regress effect size on variance (SE^2)
- Provides a bias-corrected estimate when a genuine effect exists
- Use PEESE only if PET shows significant effect

Interpretation: β_0 is the bias-corrected effect estimate

PET-PEESE in R and Stata

R code:

```
# PET
pet_model <- rma(yi, vi, mods = ~ sqrt(vi), data = dat)
print(pet_model)
# PEESE
peese_model <- rma(yi, vi, mods = ~ vi, data = dat)
print(peese_model)
```

Stata code:

```
gen se = sqrt(variance)
* PET
metareg es se [aweight=1/variance]
* PEESE
metareg es variance [aweight=1/variance]
```

What to Report: Publication Bias

Essential elements:

- Search strategy (databases, unpublished sources, grey literature)
- Comparison of published vs. unpublished effects (if available)
- Multiple bias assessment methods with results
- Interpretation of convergence/divergence across methods
- Sensitivity of main conclusions to bias corrections

Reporting and PRISMA 2020

Pre-Registration

- Pre-registering protocols for a systematic review and meta-analysis is becoming increasingly common
- Facilitates transparency and establishes a priori decisions
- Cochrane Collaboration (https://www.cochrane.org/
- Campbell Collaboration (https://www.campbellcollaboration.org
- Prospero (https://crd.york.ac.uk/prospero).
- Do it yourself using the Open Science Framework (https://osf.oi)

PRISMA 2020: What's New

PRISMA 2020 updates from 2009 version:

- Emphasis:
  - Registration and protocol
  - Risk of bias assessment
  - Certainty of evidence (GRADE)
  - Reporting of synthesis methods in detail
- PRISMA flow chart: a visual diagram that maps the number of records identified, screened, and excluded during a systematic literature review
- Separate checklists for abstracts and protocols
- More explicit guidance on reporting

Resources:

- http://www.prisma-statement.org/
- PRISMA 2020 explanation paper (BMJ, 2021)

PRISMA 2020 Flow Diagram

- Distinguish between studies and reports (one study may have multiple reports)
- Explain exclusions from synthesis

PRISMA 2020: Synthesis Reporting

- Describe methods used to synthesize results
- Specify:
  - Effect measure (e.g., OR, SMD)
  - Synthesis method (e.g., random-effects, inverse variance)
  - Estimator for τ^2 (e.g., REML, DL)
  - Handling of dependent effect sizes
  - Software used
- Methods used to assess publication bias
- Multiple methods are recommended
- Methods for assessing the certainty of evidence (e.g., GRADE)

Implementing PRISMA 2020

Practical steps:

1. **Pre-registration:** Register protocol before starting
    - PROSPERO, OSF, Cochrane, Campbell
2. **Use the checklist:** Download from PRISMA website
    - Complete during manuscript preparation
    - Submit with manuscript
3. **Create flow diagram:** Use PRISMA template
    - Track all stages of study selection
    - Be transparent about exclusions
4. **Report synthesis methods in detail:**
    - Don’t just say "random-effects meta-analysis"
    - Specify estimator, software, handling of dependencies

Alternative to PRISMA

APA (American Psychological Association) developed a reporting standard called MARS (Meta-analysis reporting standard)

Wrap Up and Q&A