p-values Had a Good Run: A Primer on the 'New Statistics'

Rob Cribbie
Quantitative Methods Program
Department of Psychology
York University

Part 7: Meta-Analysis

- Meta-Analysis
  - The statistical summarization of the effects from a set of studies investigating the same research question
  - However, the term 'meta-analysis' often also applies to the entire process of generating a research question, finding studies that investigate the research question, extracting the necessary info from the studies, and combining the results from the related studies

Systematic Review

- In some instances, “systematic review” and “meta-analysis” are used interchangeably, whereas in other instances the term *systematic review* refers to the procedures used to collect the studies of interest (i.e., those to be combined), and *meta-analysis* refers to the statistical combination of the effects from these studies
  - Systematic Review
    - A review of studies addressing a research question that is conducted according to clearly stated methods

Some History from Psychology

- 1952: Hans Eysenck concluded that there were no favorable effects of psychotherapy, starting a raging debate
  - 20 years of evaluation research and hundreds of studies failed to resolve the debate
- 1978: To prove Eysenck wrong, Gene Glass statistically aggregated the findings of 375 psychotherapy outcome studies
  - Glass concluded that psychotherapy did indeed work
- Glass called his method "meta-analysis"

The Emergence of Meta-Analysis

- The ideas behind meta-analysis predate Glass' work by several decades
- Karl Pearson (1904)
  - Averaged correlations from studies exploring the effectiveness of inoculation for typhoid fever
- R. A. Fisher (1944)
  - We can combine the results of several studies to get an appreciation for the probability associated with the aggregated data
    - Dealt primarily with combining *p*-values
- The start of the idea of *cumulating probability values,* although not specifically focused on effect sizes

The Emergence of Meta-Analysis

- W. G. Cochran (1953)
  - Discussed a method for averaging means across independent studies
  - Cochran was responsible for much of the statistical foundation for which modern meta-analysis is built upon
- Cochrane Collaboration
  - A group of researchers from around the world that conduct systematic reviews of health-care interventions and diagnostic tests and publish them in the Cochrane Library
    - e.g., https://canada.cochrane.org/

The Logic of Meta-analysis

- Traditional methods of review focus on statistical significance testing
  - E.g., the effect was statistically significant in 4 out of 7 studies
  - However, we know that NHST is highly related to sample size and not a good predictor of replication

- Meta-analysis focuses on the *direction* and *magnitude* of the effects across studies, not statistical significance
  - Direction and magnitude are represented by the effect size

When Can You Do Meta-analysis?

- Studies are empirical, not theoretical
- Results are quantitative, not qualitative
- Studies examine the same research question
- Results can be quantified in a comparable statistical form
  - i.e., effect size

Research Questions Amenable to Meta-Analysis

- Central tendency research (e.g., means)
  - Pre-post contrasts
  - Group contrasts
    - Experimentally created groups
      - E.g., comparison of treatment and control groups
    - Naturally occurring groups
      - E.g., comparing executive functioning in bilingual and monolingual individuals
- Associations among variables
  - Correlations/Regression Coefficients
    - E.g., correlation between perfectionism and depression

Answerable/Unanswerable Research Questions

- Unanswerable Research Questions
  - What is the best strategy to prevent smoking in young people?
  - How do we cure diabetes?
- Answerable Research Questions
  - Are mass media interventions effective in preventing smoking in young people?
    - E.g., smoking rates in a community from pre-intervention to post-intervention
    - Combine pre-post mean differences
  - Is sugar intake related to glycemic levels in young children?
    - Combine correlations

Which Studies to Review?

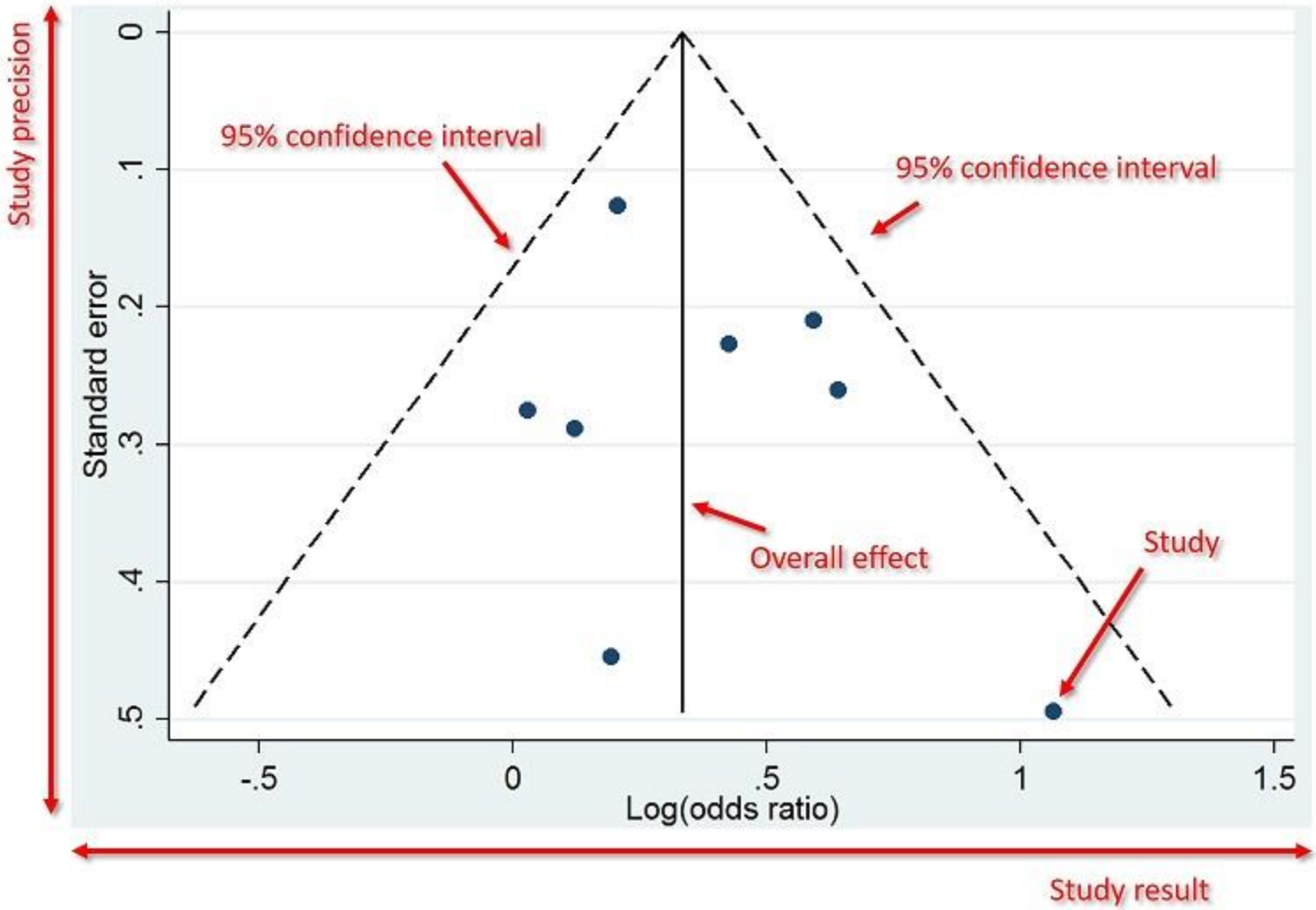
- Should be as inclusive as possible
  - Need to find ALL studies
  - Published studies are easy to find … unpublished studies are not
    - The inclusion of unpublished studies helps to minimize the effects of *publication bias*

- Apples and Oranges
  - A priori inclusion and exclusion criteria must be clear
    - It is imperative that the studies being meta-analyzed address the same research question

Exploring Publication Bias

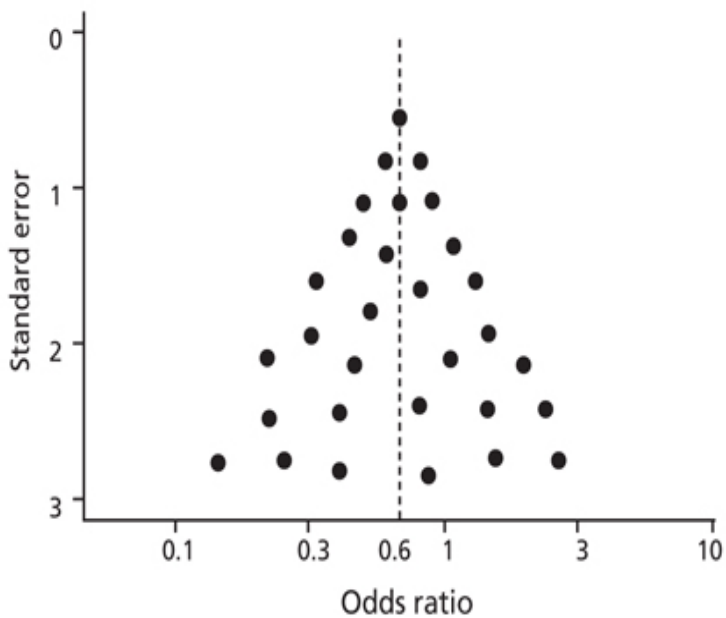
- Funnel Plot
  - A plot of the *size* of the effect of a study against the *precision* of the effect
  - Symmetrical funnel plots provide evidence of a lack of publication bias, where asymmetrical funnel plots highlight that publication bias might be present
    - E.g., if effects with low precision seem to all have larger effects, then publication bias is likely

Funnel Plot

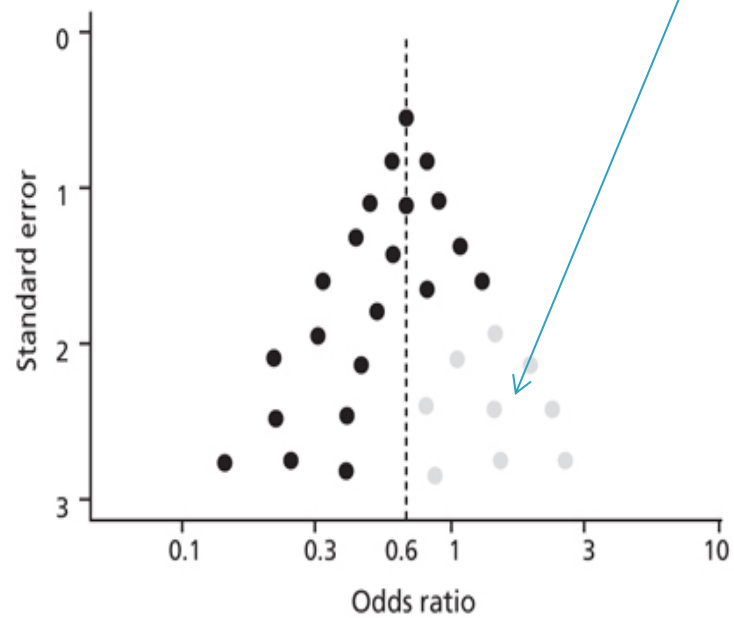

Study precision
Standard error
0
.1
.2
.3
.4
.5
95% confidence interval
95% confidence interval
Overall effect
Study
-.5
0
.5
1
1.5
Log(odds ratio)
Study result

Symmetrical vs Asymmetrical Funnel Plot

No small N studies with OR between 1 and 3



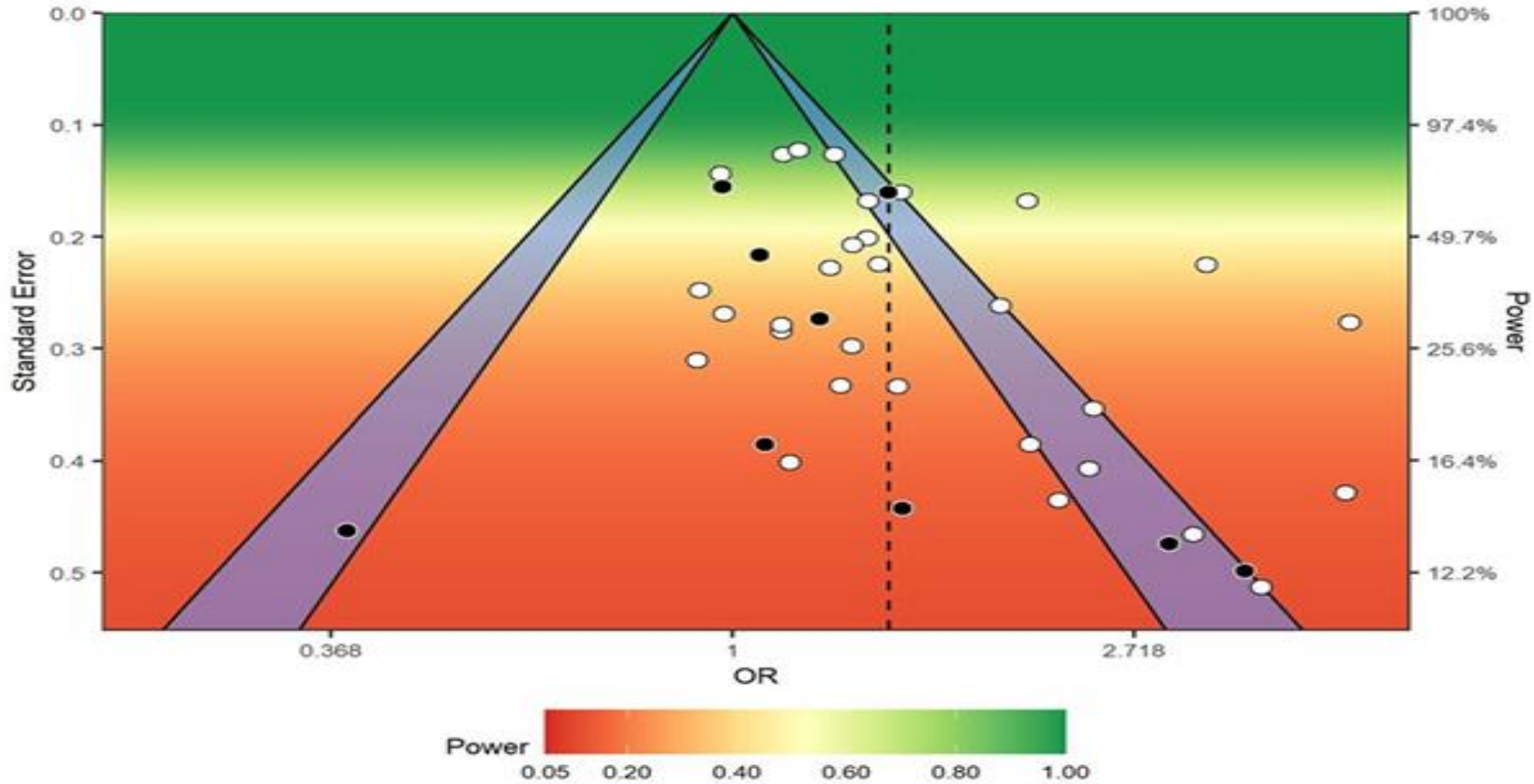
A



B

Sunset Funnel Plot

- Grey bars represent 95%/99% CIs
- White/Black dots = published/unpublished



Statistical Tests of Publication Bias

- Researchers have recommended regressing the size of the published effects on the precision of these effect (e.g., standard error) to determine if there is publication bias
- Publication bias may be present if the fitted regression model suggests that the less precise or smaller studies have bigger effect estimates than the more precise or larger studies.
- Power tends to be pretty low though for these effects, unless many studies are present

Where To Find Studies

- Computerized bibliographic databases
  - Google Scholar, PsycInfo, Medline, ERIC
- Authors working in the research domain
  - Personal websites (e.g., ResearchGate, OSF), psyarxiv
- Conference programs
- Dissertations
- Reference lists from relevant articles

What Information Should be Collected?

- Think about these long and hard before starting data collection … it is not fun to have to go back and recollect data
  - Publication details
    - Or specific location details for unpublished studies
  - Study design
  - Population details (N, characteristics)
  - Intervention/Design details
  - Operational Definitions of Variables
  - Demographics and other potential moderators
  - Outcomes
    - E.g., Means, SDs, correlations, regression coefficients, variability of coefficients, sample sizes

Why Assess the Validity of Studies?

- Lower quality studies can have biased outcome results
  - E.g., Allocation to Treatment/Control
    - Inadequate allocation concealment (e.g., investigators playing a role in allocation) exaggerated treatment effects by about 35% (Moher, 1998; Schulz, 1995)
  - E.g., Blinding
    - Lack of blinding of subjects exaggerated treatment effects by 17% (Schulz, 1995), or increased the effect size by about a half a SD (Hróbjartsson et al., 2014)

Where Can Bias be Introduced into Studies?

- Selection bias
- Allocation bias
- Confounds
- Blinding
- Data collection methods
- Withdrawals and drop-outs
- Statistical analysis
- Intervention integrity

- Summary: Lots of ways that bias can be introduced into research

Assessing the Validity of a Study

- The most common way to assess and report study quality has been using a composite, numerical scoring instrument
  - Many different quality assessment instruments are available, with most designed for randomized clinical trials
- E.g., Jadad Score for Experiments (0–3)
  - Was the study described as randomized?
  - Was the study described as double blind?
  - Was there a description of withdrawals and dropouts?

Methodological Quality Dilemma

- Include or exclude low quality studies?
  - The findings of all studies are potentially in error (methodological quality is a continuum, not a dichotomy)
    - Being too restrictive may limit ability to generalize
    - Being too inclusive may weaken the confidence that can be placed in the findings
  - Methodological quality is often subjective
    - You must strike a balance that is appropriate to your research question
- When including low quality studies, you can weigh effects by study quality or explore study quality as a moderator

Level of Replication

- Replications can range from “conceptual” replications to “pure” or “direct” replications
  - Direct replications are the repetition of an experimental procedure to as exact a degree as possible, whereas a conceptual replication is the use of different methods/procedures to repeat the test of a hypothesis
- You must be able to argue that the collection of studies you are meta-analyzing examine the same relationship
- The closer to pure replications your collection of studies, the easier it is to argue comparability of the effects from each study

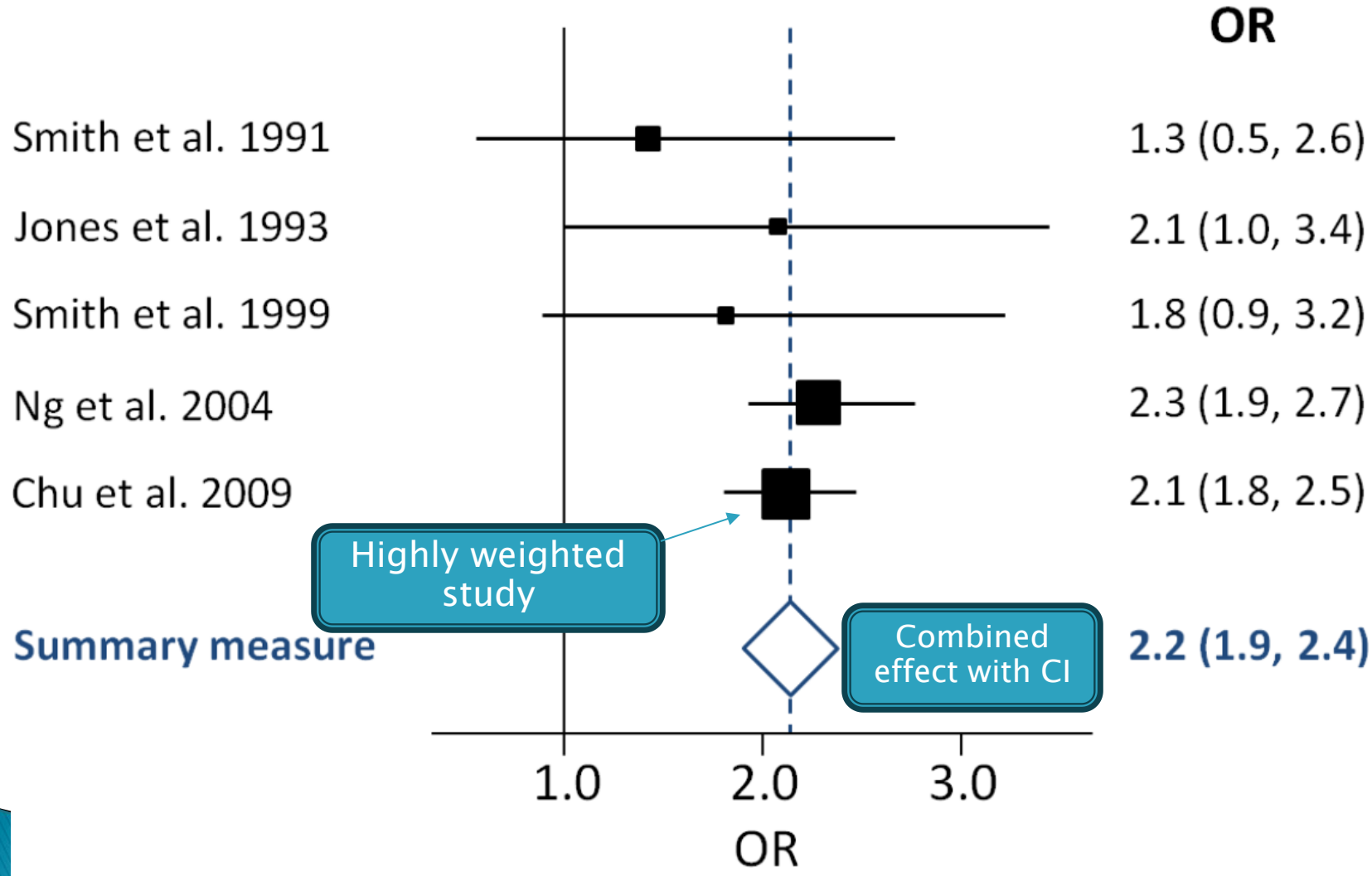
Effect Size in Meta-Analysis

- Effect size is the "dependent variable"
  - Since studies usually differ in the nature/scale of outcome/predictor variables, standardized effect size measures are almost always used
    - I.e., studies must be able to be directly compared
  - A standardized index must be comparable across studies, represent the magnitude and direction of the relationship of interest, and be independent of sample size
    - E.g., standardized mean difference (e.g., Cohen's *d*), correlation coefficient (e.g., Pearson's *r*) , odds-ratio
- We discussed effect sizes in detail

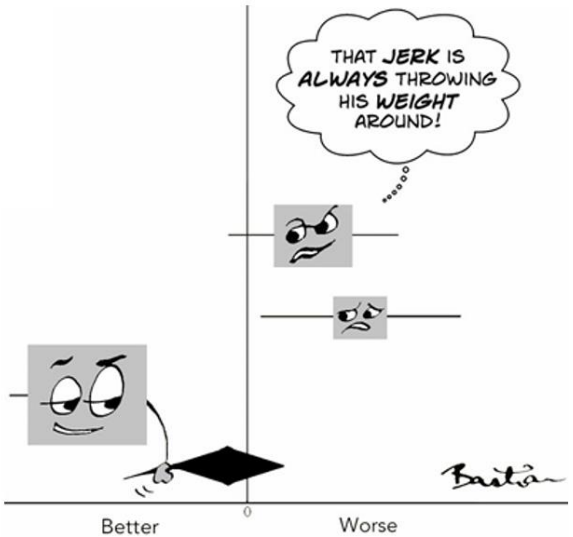
Forest Plot

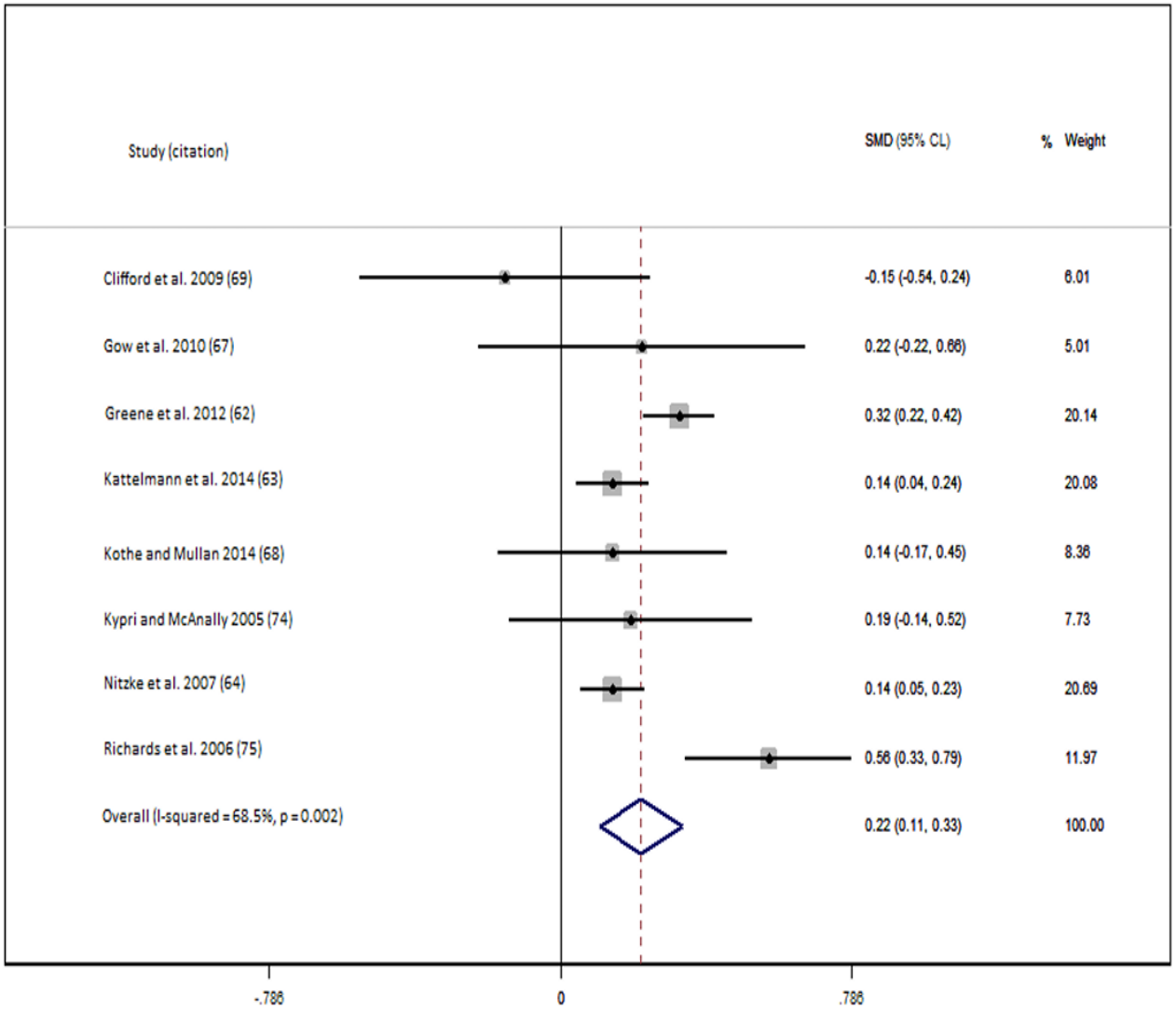
- A visual representation of the effect sizes (and confidence intervals for the effect sizes) of the multiple studies included in a systematic review
  - Reminder: all effects must be measured in the same metric, e.g., correlation
- The size of the effect size icons (e.g., squares) indicates the "weight" of the study to the combined effect
  - E.g., larger N studies have a higher weight
- The plot also shows the effect size (and confidence interval for the effect size) of the combined effect across studies

Forest Plot Example – Odds Ratios



Forest Plot Example – Cohen's *d*





Fixed Effects vs Random Effects

- There are two popular statistical models available for conducting a meta-analysis
  - In other words, two models available for arriving at a "combined" measure of effect size
  - Fixed Effects Model
    - Assumes that all the studies investigated the same population, and therefore estimate the same population effect size
      - Highly questionable
  - Random Effects Model
    - Allows for the possibility that the studies investigated somewhat different populations, and therefore estimate different population effect sizes

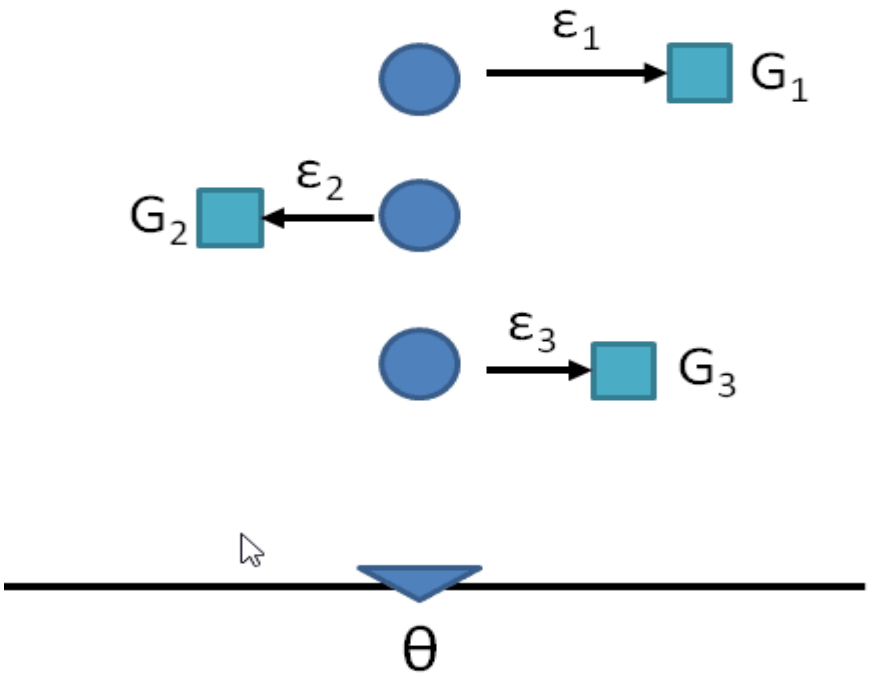
Fixed Effects vs Random Effects

- It is difficult to imagine a setting in which multiple studies conducted in different locations, with different samples, and with potentially different measures, are all studying the same population (and thus after a single population effect size)
- The random effects model is more realistic and provides a basis for understanding the heterogeneity of effect sizes
  - Further, the models give the same answer if there is only a single population, so it is hard to find a reason for a researcher to prefer a fixed effects model

Fixed Effects Meta-Analysis

- We assume that each observed study effect size (G_i) is an estimate of a fixed effect size, θ
- The difference between G_i and θ is sampling error (ϵ_i)

$$G_i = \theta + \varepsilon_i$$



Fixed Effects Meta-Analysis

- For a set of S effect size measures (γ)
  - $\widehat{\gamma}_F = \frac{\sum_{i=1}^{S} w_i \widehat{\gamma}_i}{\sum_{i=i}^{S} w_i}$
  - $w_i = \frac{1}{s^2(\widehat{\gamma}_i)} = \frac{1}{SE(\widehat{\gamma}_i)^2}$
  - $s^2(\widehat{\gamma}_F) = \frac{1}{\sum_{i=i}^{S} w_i}$

We are switching from G to γ (more common in the literature)

Fixed Effects Meta-Analysis: Example

- Study 1: $M_1 = 12$, $M_2 = 14$, $SD_1 = 3$, $SD_2 = 3$, $n_1 = 22$, $n_2 = 32$
- Study 2: $M_1 = 14$, $M_2 = 16$, $SD_1 = 2$, $SD_2 = 2$, $n_1 = 25$, $n_2 = 52$
- Study 3: $M_1 = 11$, $M_2 = 13$, $SD_1 = 4$, $SD_2 = 4$, $n_1 = 142$, $n_2 = 128$

Cohen's *d* Values

$$d = \frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_s^2}{n_1 + n_2 - 2}}}$$

$$d_1 = \frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_s^2}{n_1 + n_2 - 2}}} = \frac{12 - 14}{\sqrt{\frac{(22 - 1)3^2 + (32 - 1)3^2}{22 + 32 - 2}}} = -.67$$

Study 1 Cohen's *d*

$$d_2 = \frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_s^2}{n_1 + n_2 - 2}}} = \frac{14 - 16}{\sqrt{\frac{(25 - 1)2^2 + (52 - 1)2^2}{25 + 52 - 2}}} = -1.00$$

Study 2 Cohen's *d*

$$d_3 = \frac{M_1 - M_2}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_s^2}{n_1 + n_2 - 2}}} = \frac{11 - 13}{\sqrt{\frac{(142 - 1)4^2 + (128 - 1)4^2}{142 + 128 - 2}}} = -.50$$

Study 3 Cohen's *d*

Fixed Effects Meta-Analysis: Example

- Study 1: $M_1 = 12$, $M_2 = 14$, $SD_1 = 3$, $SD_2 = 3$, $n_1 = 22$, $n_2 = 32$
- Study 2: $M_1 = 14$, $M_2 = 16$, $SD_1 = 2$, $SD_2 = 2$, $n_1 = 25$, $n_2 = 52$
- Study 3: $M_1 = 11$, $M_2 = 13$, $SD_1 = 4$, $SD_2 = 4$, $n_1 = 142$, $n_2 = 128$

Variances of the *d* values

$$s^2(d) = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)}$$

$$s^2(d_1) = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)} = \frac{22 + 32}{(22)(32)} + \frac{-.67^2}{2(22 + 32 - 2)} = .085$$

$$s^2(d_2) = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)} = \frac{25 + 52}{(25)(52)} + \frac{-1.00^2}{2(25 + 52 - 2)} = .073$$

$$s^2(d_3) = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)} = \frac{142 + 128}{(142)(128)} + \frac{-.50^2}{2(142 + 128 - 2)} = .016$$

The study with the largest N has the smallest s^2

Study 1 variance

Study 2 variance

Study 3 variance

Fixed Effects Meta-Analysis: Example

- Study 1: $M_1 = 12$, $M_2 = 14$, $SD_1 = 3$, $SD_2 = 3$, $n_1 = 22$, $n_2 = 32$
- Study 2: $M_1 = 14$, $M_2 = 16$, $SD_1 = 2$, $SD_2 = 2$, $n_1 = 25$, $n_2 = 52$
- Study 3: $M_1 = 11$, $M_2 = 13$, $SD_1 = 4$, $SD_2 = 4$, $n_1 = 142$, $n_2 = 128$

Weights

$$w = \frac{1}{s^2(d)}$$

$$w_1 = \frac{1}{s^2(d)} = \frac{1}{.085} = 11.73$$

$$w_2 = \frac{1}{s^2(d)} = \frac{1}{.073} = 13.78$$

$$w_3 = \frac{1}{s^2(d)} = \frac{1}{.016} = 63.34$$

The study with the smallest s^2 has the largest weight

Study 1 weight

Study 2 weight

Study 3 weight

Fixed Effects Meta-Analysis: Example

$$\hat{\gamma}_F = \frac{\sum_{i=1}^{S} w_i \hat{\gamma}_i}{\sum_{i=i}^{S} w_i} = \frac{\overset{\textbf{Weight 1}}{(11.73)}\overset{\textbf{d1}}{(-.67)}+\overset{\textbf{Weight 2}}{(13.78)}\overset{\textbf{d2}}{(-1.00)}+\overset{\textbf{Weight 3}}{(63.34)}\overset{\textbf{d3}}{(-.5)}}{11.73+13.78+63.34} = -.60$$

Mean Effect Size

$$s^2(\hat{\gamma}_F) = \frac{1}{\sum_{i=i}^{S} w_i} = \frac{1}{11.73+13.78+63.34} = .011$$

Variance of Combined Effect size Estimate

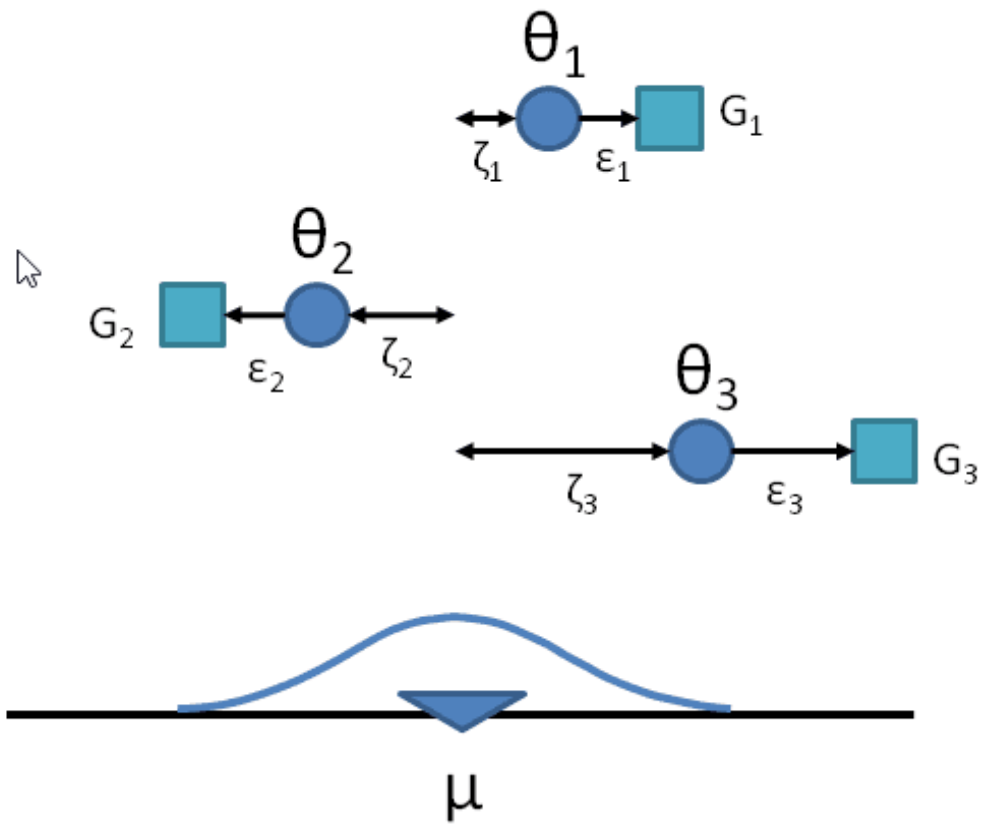
$$\mathrm{SE}(\hat{\gamma}_F) = \sqrt{s^2(\hat{\gamma}_F)} = \sqrt{.011} = .10$$

$$95\%\ \mathrm{CI}(\hat{\gamma}_F) = \hat{\gamma}_F \pm (1.96)\mathrm{SE}(\hat{\gamma}_F) =$$
$$\{(-.60 - 1.96 * .10), (-.60 + 1.96 * .10)\} = \{-.80, -.40\}$$

Confidence Interval

Random Effects Meta-Analysis

- In addition to sampling error (ϵ_i), in random effects model we also have true variation in effect sizes (ζ_i)



Random Effects Meta-Analysis

- For a set of *S* effect size measures (γ)
  - $\hat{\gamma}_R = \frac{\sum_{i=1}^{S} w_i \hat{\gamma}_i}{\sum_{i=i}^{S} w_i}$
  - $w_i = \frac{1}{s^2(\hat{\gamma}_i)+\tau^2}$
  - $\tau^2 = \frac{Q-(S-1)}{\sum_{i=i}^{S} w_i - \frac{\sum_{i=i}^{S} w_i^2}{\sum_{i=i}^{S} w_i}}$ (for $Q > S-1$)
  - $Q = \sum_{i=i}^{S} w_i \left(\hat{\gamma}_i - \hat{\gamma}_F\right)^2$

Same formula, but different components, as $\hat{\gamma}_F$

τ^2/Q estimate the dispersion of the individual effects around the fixed effect (i.e., study heterogeneity

Heterogeneity of Effect Sizes

- A simple goodness-of-fit test can be used to test for excessive heterogeneity
  - Q $\sim \chi^2_{df=S-1}$
  - We reject the null that there is no population heterogeneity if Q $\geq \chi^2_{\alpha,\, df=S-1}$

- The problem with this approach is that the test has low-power when *S* is small

Proportion of Variability due to Study Heterogeneity

- A better approach to quantifying heterogeneity is to use an effect size measure
- $I^2 = \frac{Q-S+1}{Q}$
- I^2 ranges from 0 to 1, with larger values indicating more heterogeneity
- Represents the proportion of total variability in the effect estimates that is due to heterogeneity rather than sampling error (chance)

Proportion of Variability due to Study Heterogeneity

- Interpretation
  - $I^2 = 0\%$ indicates no observed heterogeneity (all variability in effect sizes is due to sampling error)
  - 0% to 40%: Low heterogeneity
  - 40% to 80%: Moderate/substantial heterogeneity
  - 80% to 100%: Considerable heterogeneity
    - If I^2 is large, it makes sense to investigate potential sources of heterogeneity (e.g., meta-regression)
- Limitations of I^2:
  - I^2 is influenced by the precision of the included studies
    - With highly precise studies (small CIs), even small differences in effect sizes can lead to high I^2 values
  - I^2 can be unstable when the number of studies is small, leading to imprecise estimates.

Summary: Steps of a Systematic Review/Meta-Analysis

- Specify your research question/effect of interest
- Find studies that investigate the effect of interest
- Extract all necessary information from the studies
- Assess the validity of the studies and determine inclusion/exclusion/weighting
- Estimate the combined effect size and CI for the effect size
- Explore moderators of the variability in effect sizes
- Interpret the findings

Strengths of Meta-Analysis

- Imposes strict procedures on the process of summing up research findings
- Can handle a large numbers of studies, which would be difficult in a qualitative review
- Represents findings in a more sophisticated manner than conventional reviews
- Capable of detecting moderators of effects

Weaknesses of Meta-Analysis

- Requires a lot of effort!
- Mechanical aspects don't lend themselves to capturing more qualitative distinctions between studies
- "Apples and oranges"
  - Comparability of studies is often questionable
- Most meta-analyses include "blemished" studies
- Selection bias possesses continual threat
  - E.g., Null finding studies are hard to find