Using Meta-Analysis in Evaluation Research: A Demonstration

American Evaluation Association

Ning Rui, Ph.D.

Kathryn Henderson, Ph.D.

October 17, 2013

Overview

- What is meta-analysis and why would you use it?
- What are the advantages and limitations of meta-analysis?
- What are the necessary steps for completing a meta-analysis?
- A demonstration

What is Meta-Analysis?

Meta-analysis is a method of conducting a systematic review of the literature, combining results from all relevant studies in the hopes of identifying patterns among the study results.

Why Use Meta-Analysis?

Meta-analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature. (Glass, 1976)

When To Use Meta-Analysis?

- To establish the presence of an effect.
- To determine the magnitude of an effect.
- To resolve differences in a literature.
- To determine important moderators of an effect.

What are Common Research Areas for Meta-Analysis?

- Epidemiological research
- Education research
- Labor market research
- Criminal justice research

Advantages of Meta-Analysis

- Meta-analysis is able to integrate findings from a large number of studies of a given topic.
- Results can be generalized to a larger population.
- Use of more data increases the precision and accuracy of estimates.
- Inconsistency of results across studies can be quantified and analyzed.
- Moderators can be included to explain variation between studies.

Limitations of Meta-Analysis

Publication bias

A meta-analysis reflects only what's published.

Heterogeneity bias

Rigorous meta-analysis cannot handle too much heterogeneity in the study design or program implementation.

‘Weak link’ bias

A meta-analysis is only as good as the studies it summarizes.

Utility of Meta-Analysis for Evaluation

Meta-analysis:

- Reveals gaps in knowledge about programs.
- Allows for analysis of results across multiple programs/sites/time periods.
- Shows the impact of programs on different populations.
- Minimizes selection bias and error.

How to do Meta-Analysis

1. Define the theoretical relationship of interest.
2. Determine selection criteria of studies for inclusion in analysis.
3. Collect the population of studies that provide data on the relationship.

How to do Meta-Analysis (cont.)

4. Determine the outcome variables and measures.
5. Extract and calculate effect sizes.
6. Select appropriate statistical models.
7. Examine the distribution of effect sizes and analyze the impact of moderating variables.
8. Interpret and report the results.

Determining Outcome Measures

- Are there operational definitions of the outcomes?
- Are there measurable constructs that can address your research questions? (e.g., Academic achievement: GPA, test scores, homework scores, grades, grade promotion, graduation)
- Is the outcome continuous or discrete variable?
- Are there multiple outcomes (primary vs. secondary)?
- Is the construct measured by standardized instruments?
- Are measures across different studies based on the same scale?

Calculation of Effect Size Estimates

- Continuous Outcome: Standardized mean difference (Cohen's *d* or Hedge's *g*)
  - Group contrast (e.g., treatment versus control)
  - Inherently continuous outcome construct
- Discrete Outcome: Odds-ratio and Risk-ratio (*OR* and *RR*)
  - Group contrast (e.g., treatment versus control)
  - Inherently dichotomous (binary) outcome construct
- Correlation coefficient (*r*)

Calculation of Effect Size Estimates

Formula for *Cohen's d:*

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

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

d can be converted from *t:*

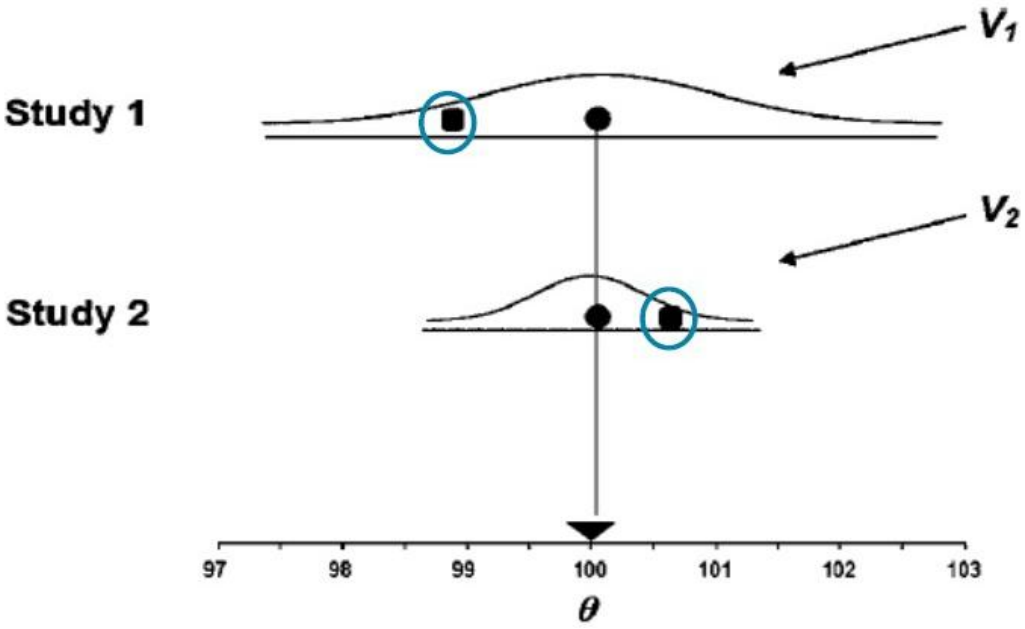
$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$$

Binary outcome Odds Ratio:

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

Statistical Models

- Fixed-Effect Model: Assume that the effect size estimates across studies share a common mean and differ due to sampling error only

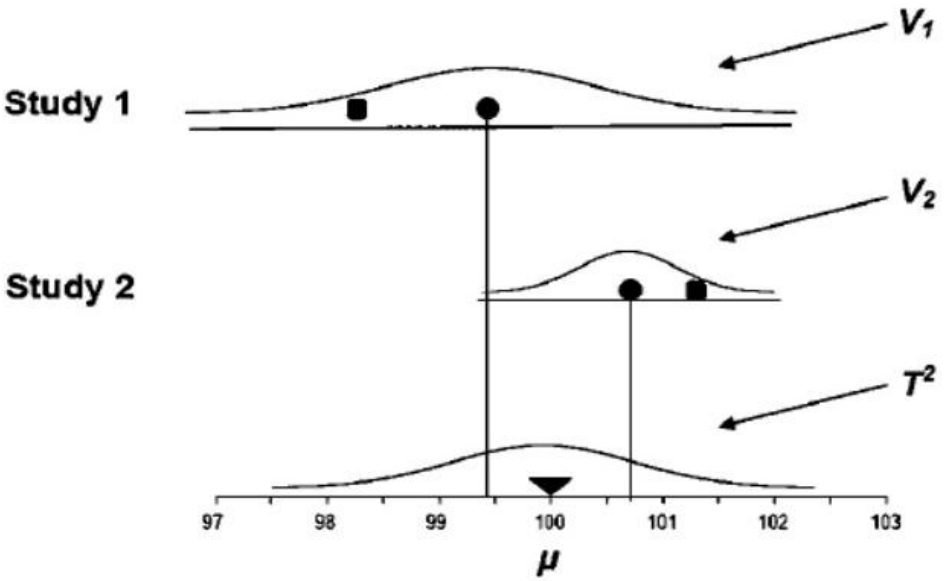


Statistical Models

- Random Effects Model:

-There are multiple population effects; ES variation is due to sampling error & variation in population of effects

-REM will estimate the mean and variance of the underlying population



Statistics to Measure Heterogeneity

- *Q* statistic: a weighted sum of squares of difference between individual ES and the mean ES

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

- *I* squared: adjusted by degrees of freedom when the number of studies is low and the sample sizes in individual studies are small

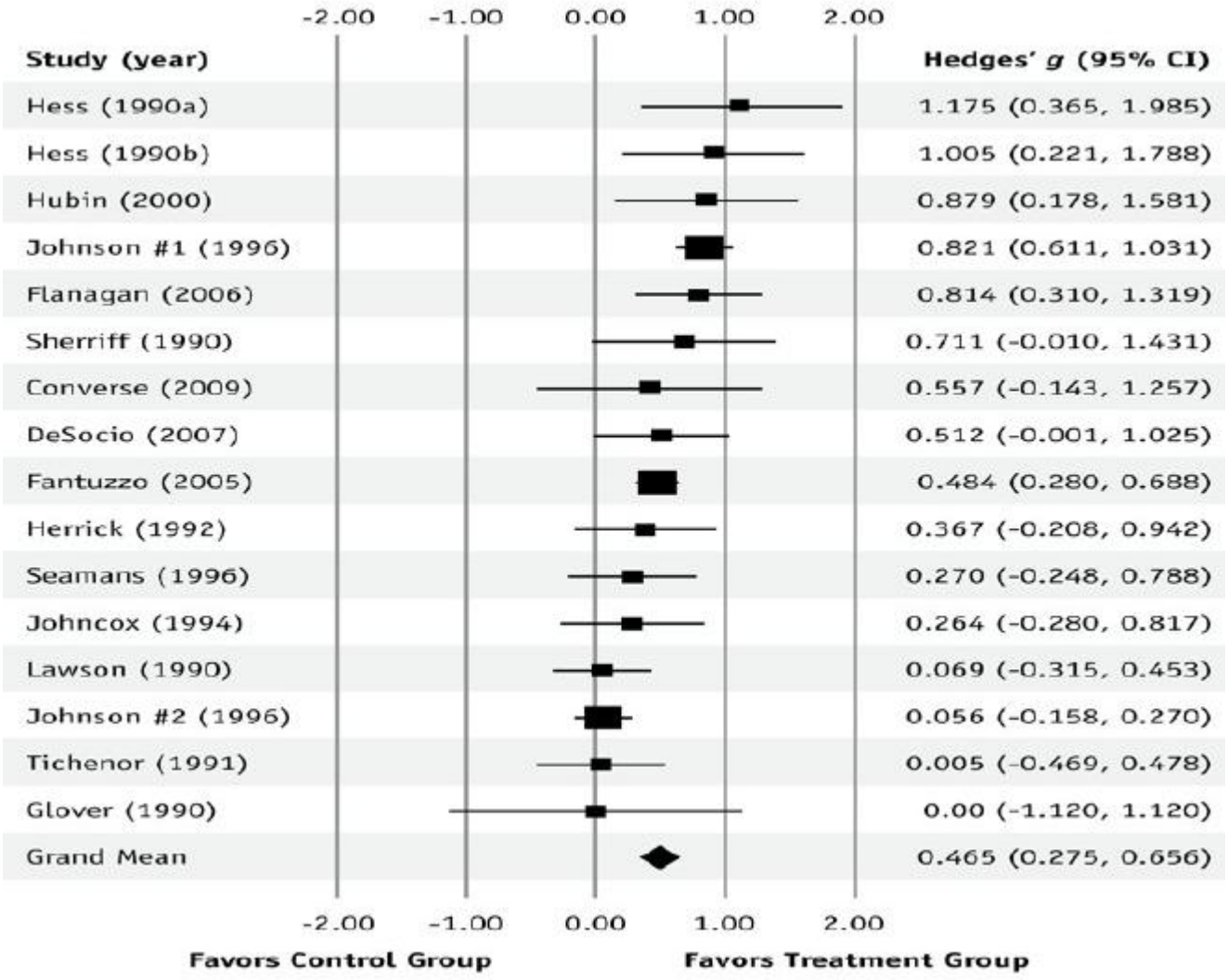
$$I^2 = 100\% \times \frac{Q - df}{Q}$$

- When *I* squared is above 90%, the fixed-effect model results are unreliable

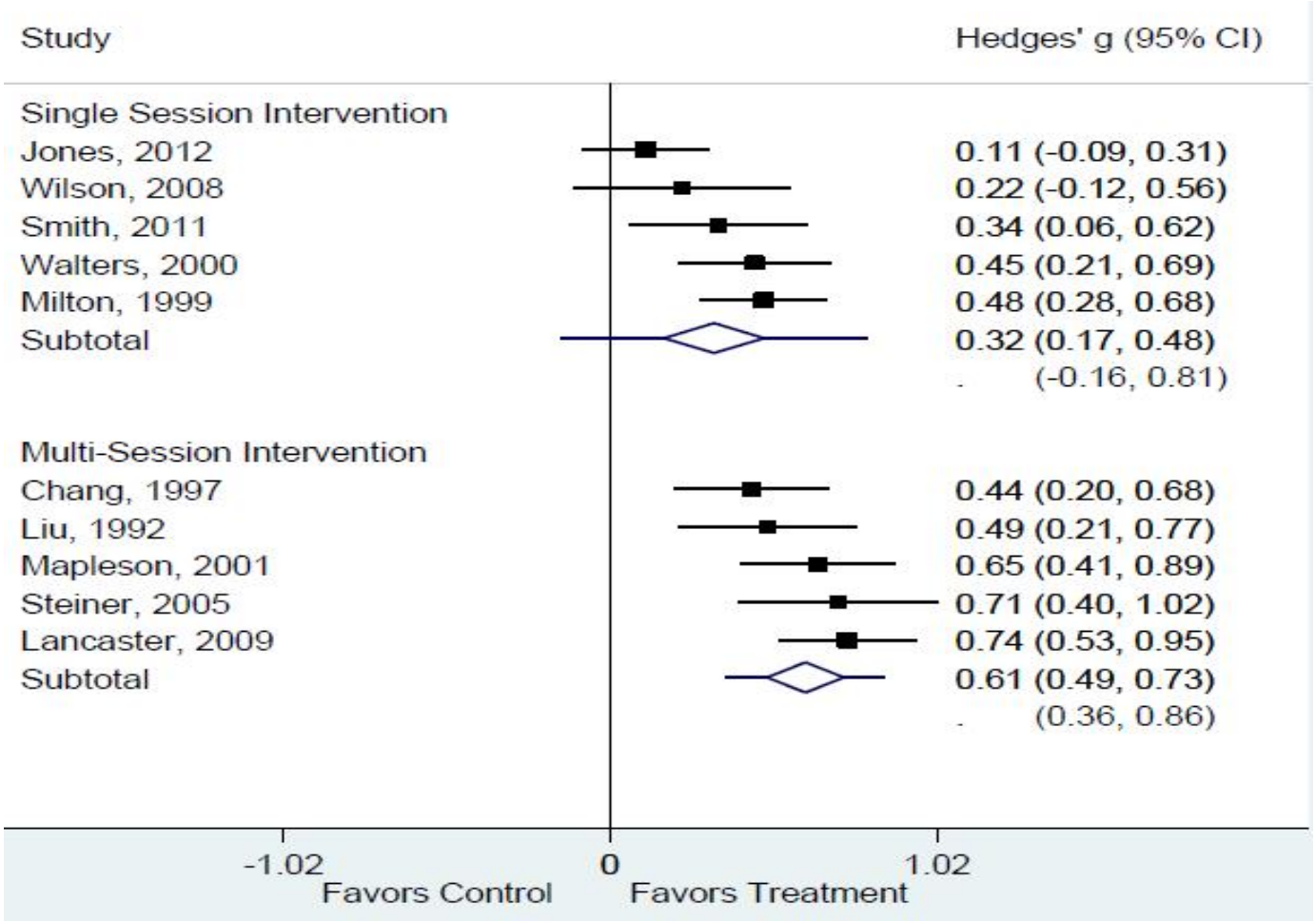
Graphic Display of Results

- Goal: to convey the results of meta-analysis to readers
- Things to pay attention:
  - Titles, captions, legends
  - Line thickness, symbol size (usually weighted by N)
  - Reasonable range of scale
- Types of Graphs
  - Forest plot (most common)
  - Funnel plot
  - Bubble plot

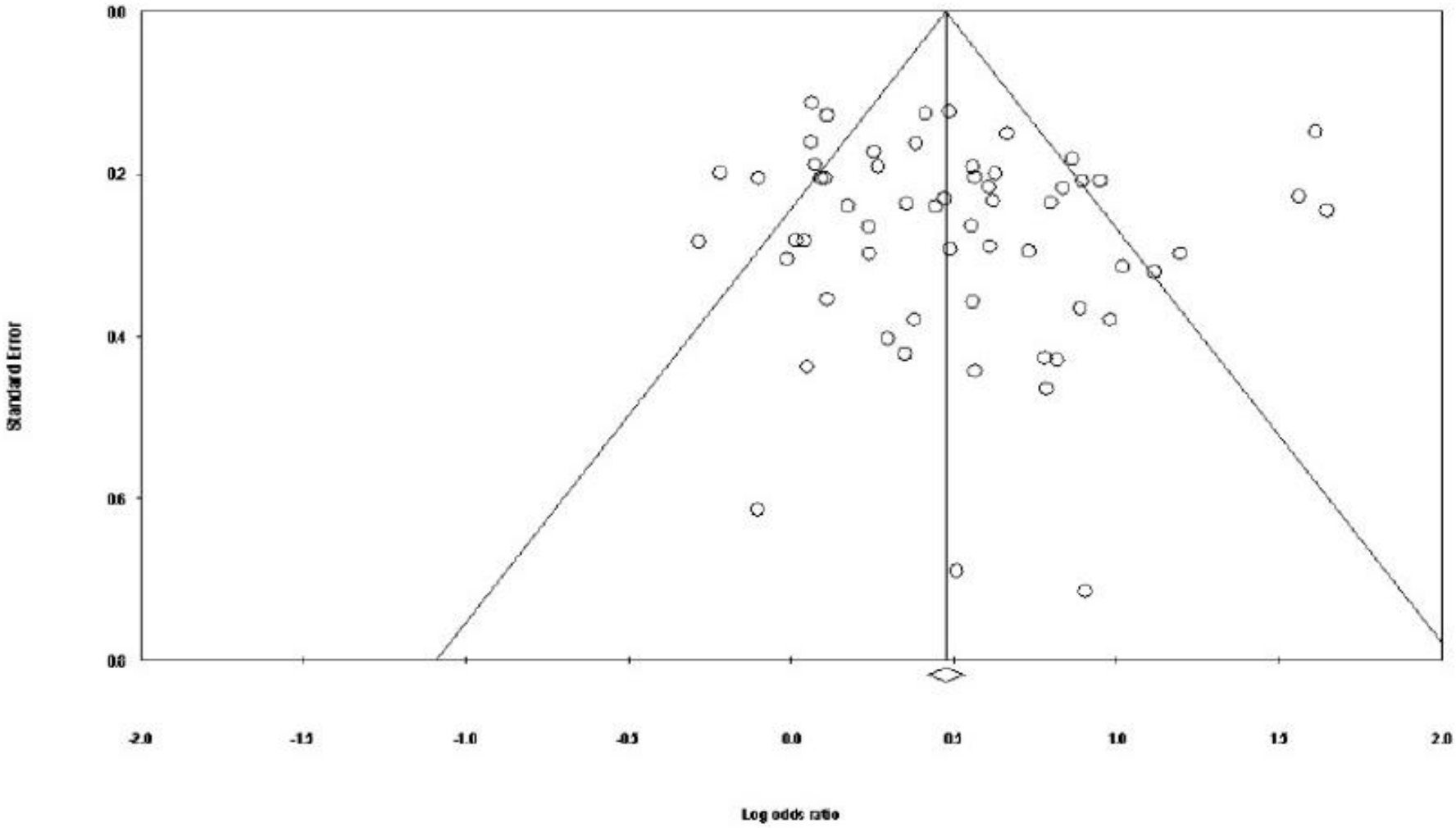
Graphic Display of Results (Forest plot)



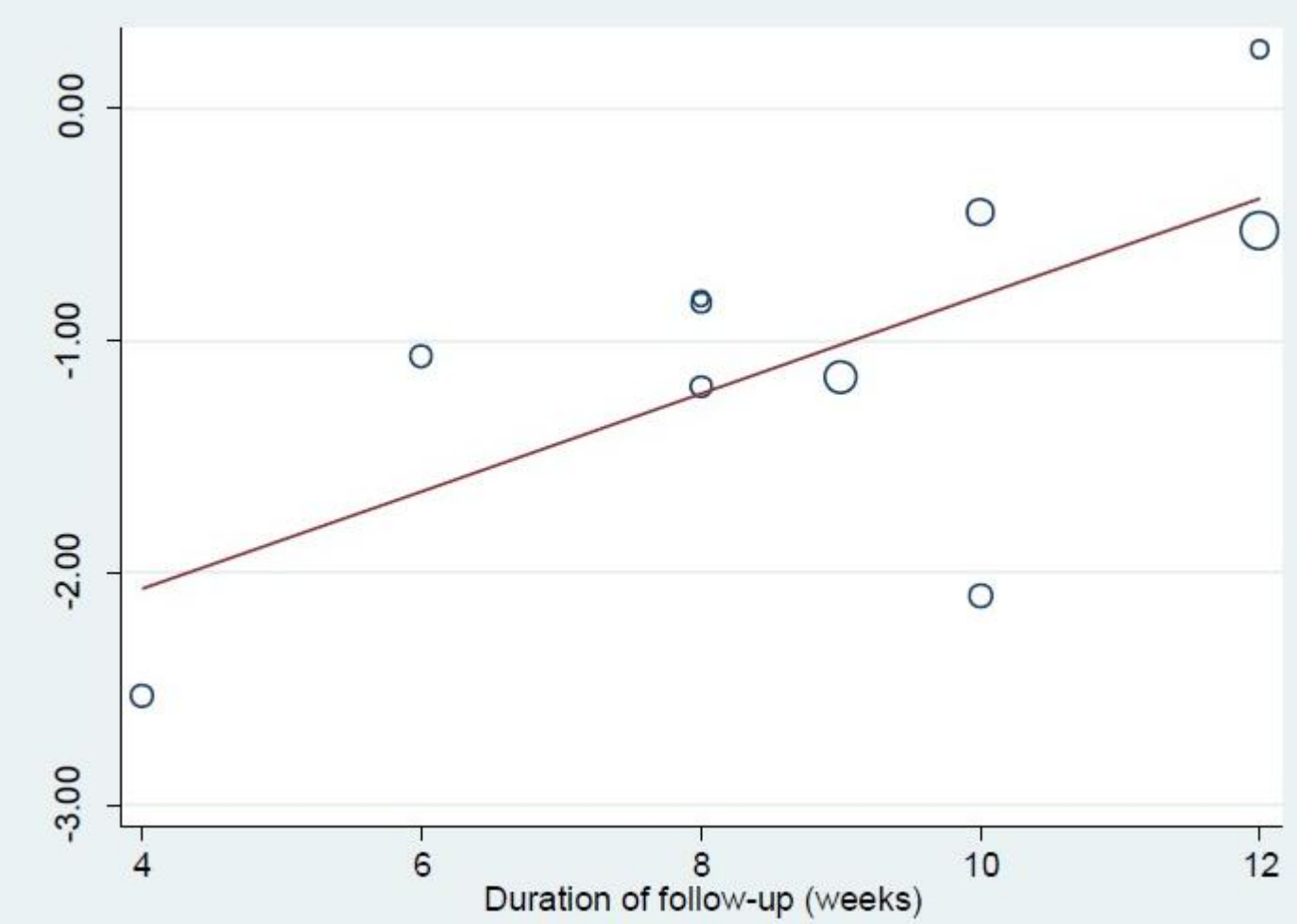
Graphic Display of Results (Forest plot with subgroup)



Graphic Display of Results (Funnel plot)



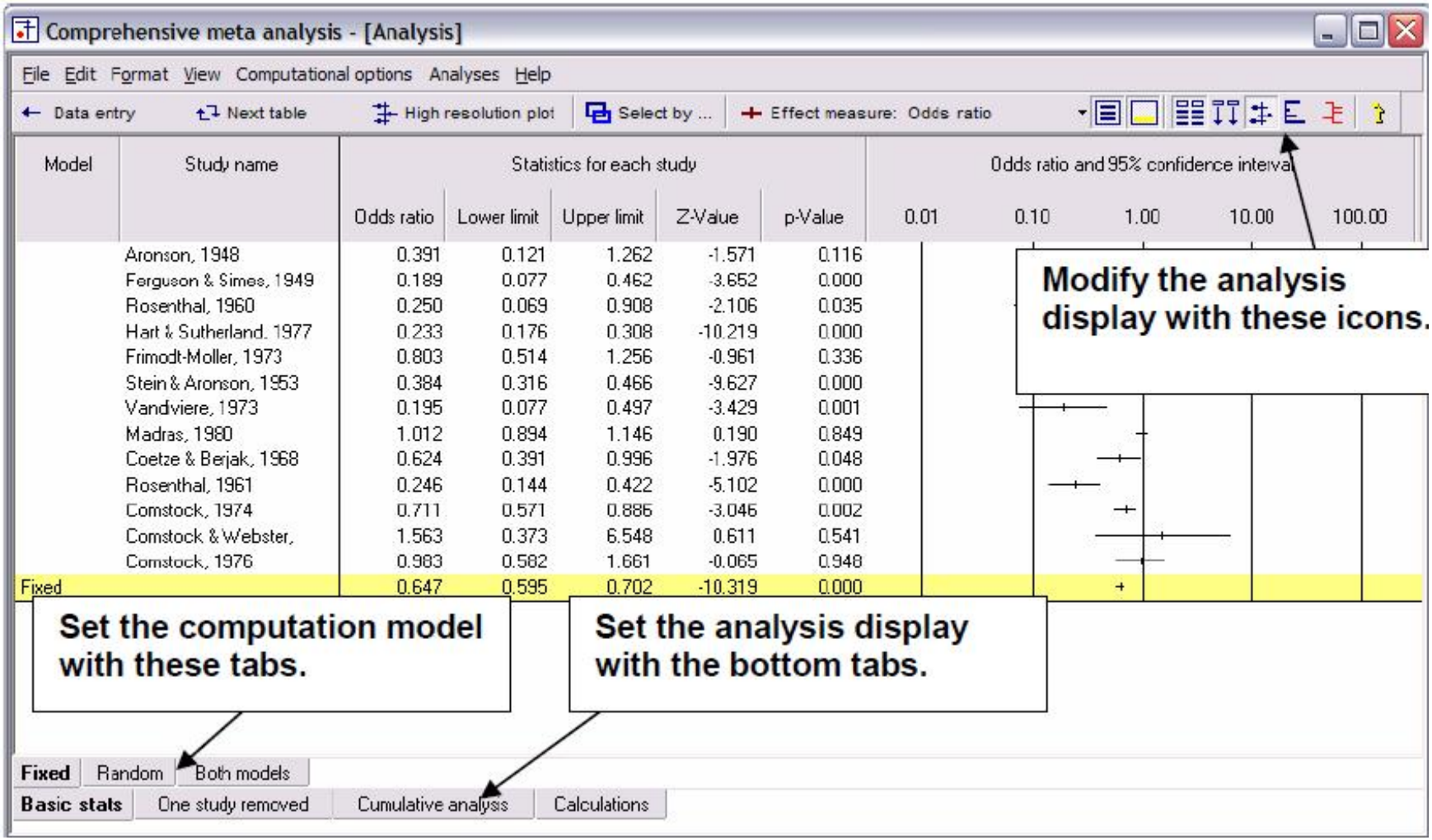
Graphic Display of Results (Bubble plot)



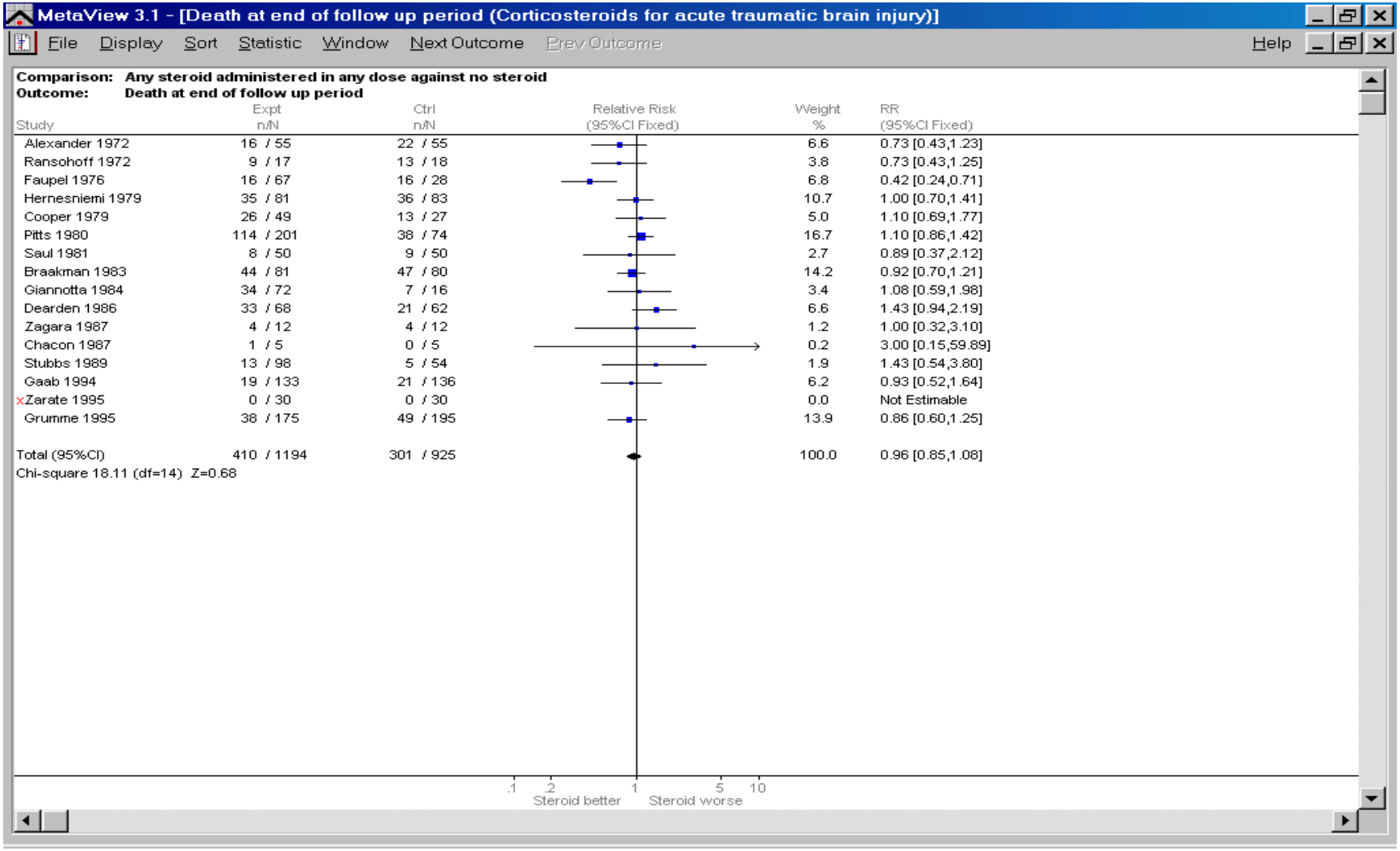
Software Options

- Comprehensive Meta-Analysis (CMA) (commercial)
- RevMan (free)
- MIX (free statistical add-in in Excel)
- SPSS, STATA and SAS Macros (downloadable from David Wilson's website: http://mason.gmu.edu/~dwilsonb/ma.html)

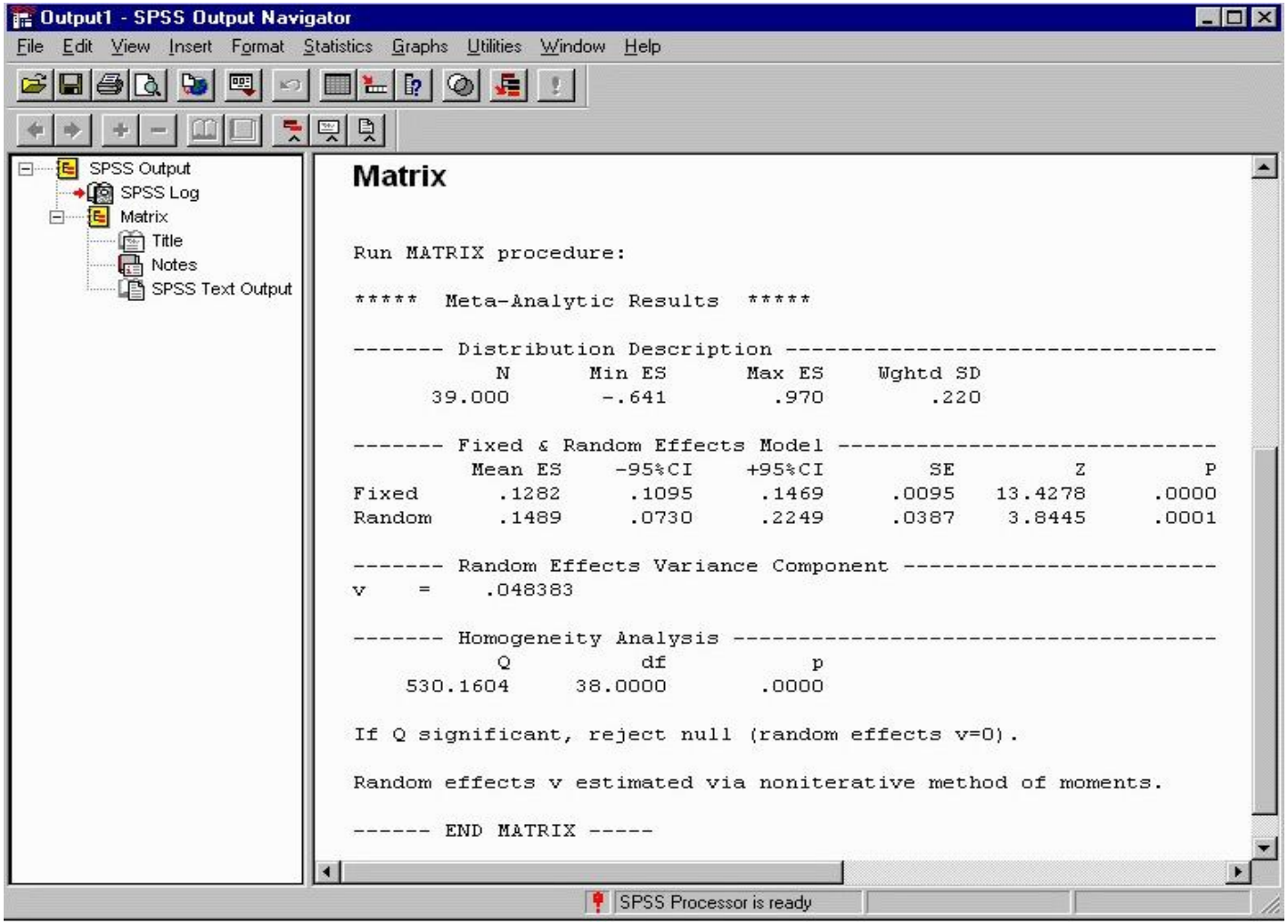
Screenshot of Output Using CMA



Screenshot of Output Using RevMan



Screenshots of Output Using SPSS Macro



Demonstration

Objective

- To review and synthesize evidence about academic and non-academic effects of classroom detracking.

Research Questions

- Are there differences in academic outcomes between average and high achieving students and low achieving students?
- What are the non-academic outcomes of detracking for all students in K-12 schools?

Study Selection Criteria

1. Comprehensive detracking with most students included; excluded programs for gifted, special education
2. Restricted to North America and written in English
3. Must clearly describe a practice that can be identified as detracking or heterogeneous grouping
4. N ≥ 30 students
5. Duration ≥ 1 semester

Preference was given to experimental studies compared to observational studies

Commonly Used Search Sources

- PubMed
- PsychInfo
- Education Resources Information Center (ERIC)
- JSTOR
- Public Affairs Information Service (PAIS)
- Web of Science/Web of Knowledge
- Pro quest digital dissertation databases
- GoogleScholar
- University libraries
- Reference lists

Literature Search and Review

- *Sources*: University libraries, Education recourses Information Center (Eric), PsychINFO, Pro quest digital dissertation databases
- *Keywords*: detracking; school integration; heterogeneous grouping; tracking; ability grouping; or homogenous grouping
- *Studies included*: Fifteen studies conducted from 1972 to 2006 were located and reviewed, including 4 experimental studies, 2 quasi-experimental studies, 7 observational studies, and 2 qualitative studies.

Effect Size Calculation

- Detracking effect size (Cohen's *d*) computed for each ability group in each study, then compared using **forest plot**
- A pooled detracking effect estimate as a weighted average of the ES computed from individual studies with n as weight

$$Weighted\ average = \frac{sum\ of\ (effect \times weight)}{sum\ of\ weights} = \frac{\Sigma TiWi}{\Sigma Wi}$$

Model Selection

- Both fixed-effects and random-effects models were conducted for all and subsets of studies (by the academic ability of students and research design)

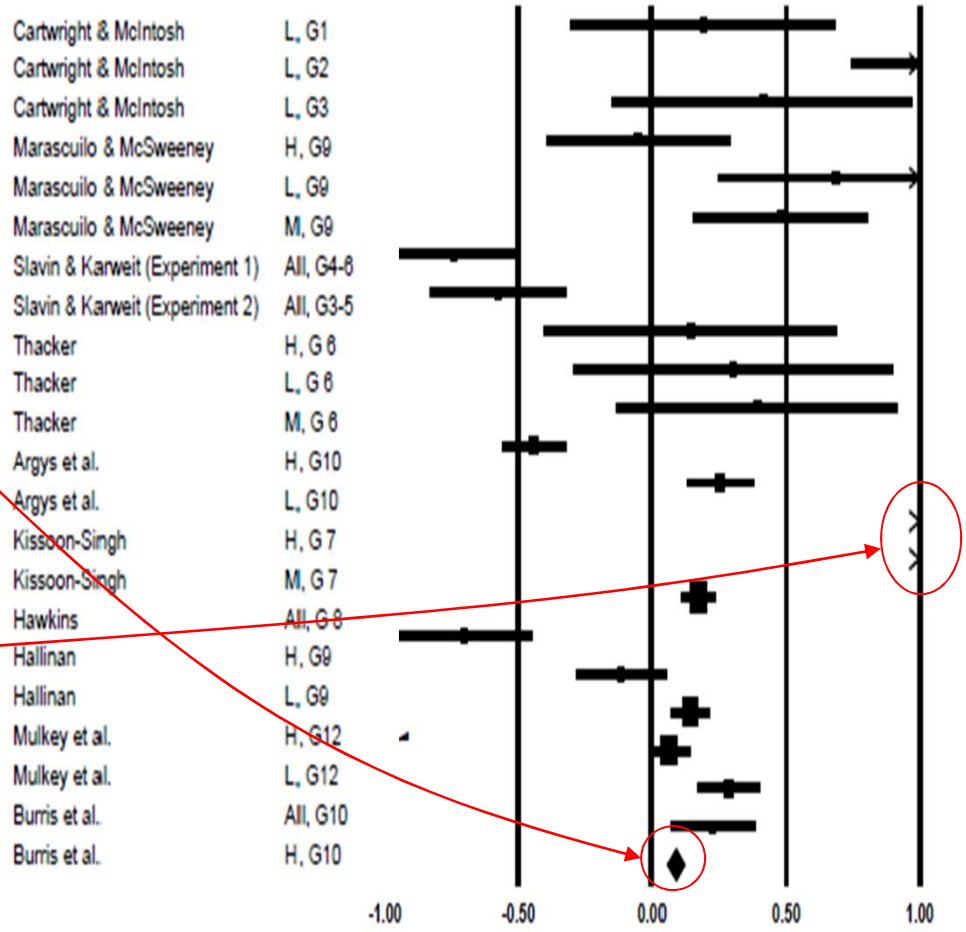
- I^2 index is used to assess heterogeneity in findings across studies (Higgins and Thompson, 2002)

Interpreting the Results (Overall Impact)

- Overall, students in detracked groups performed slightly better academically (*k*= 22, *N*= 15,577)

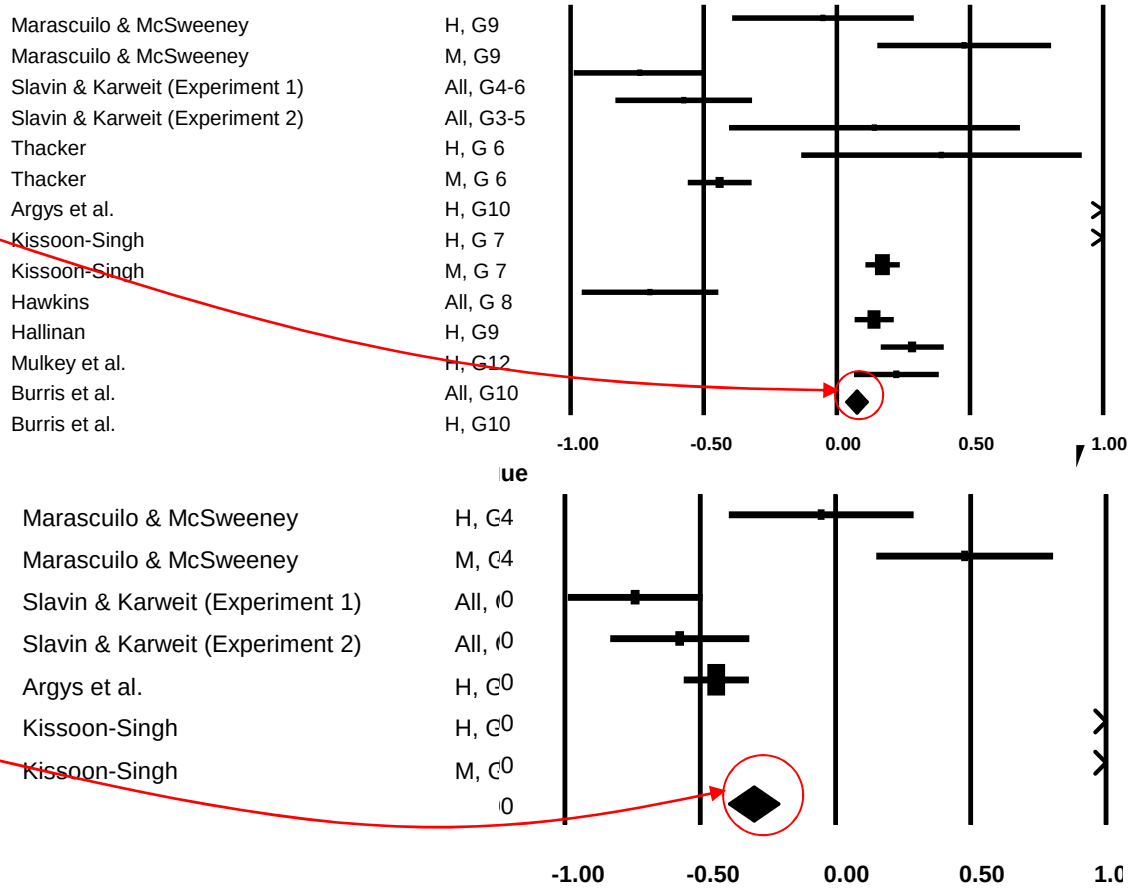
  Fixed: ***d*= .09**, *p* <.0001
  Random: ***d*= .20**, *p* <.01

- Results are highly heterogeneous with I^2 *(21)* = 94%; one study is off the chart with extremely high ES



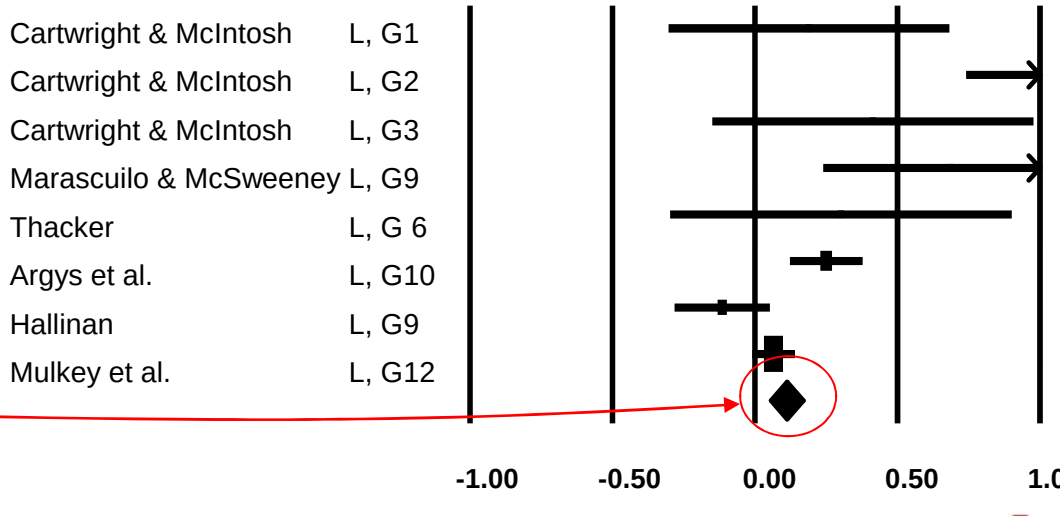
Moderator Analysis (Average/High Ability)

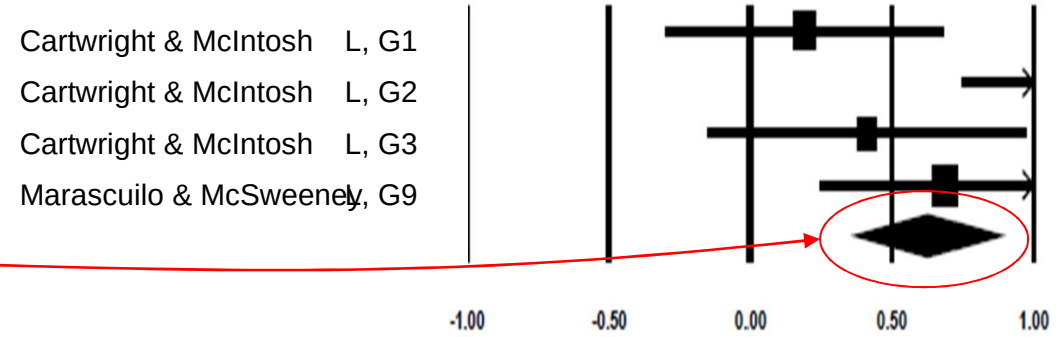
- Fixed: Detracked groups performed slightly better (d = 0.08, k = 14, p <.05), but I^2 *(14)* > 95%, making the results unreliable
- Random: No impact on average- to high-ability students [p =.13, 95% CI (-.047, .388)].
- When only RCTs included, overall effect becomes negative (d = -0.30, k = 7, p < 0.0001), but random-effects model derived non-statistically significant outcome (p= 0.125).



Moderator Analysis (Lower Ability)

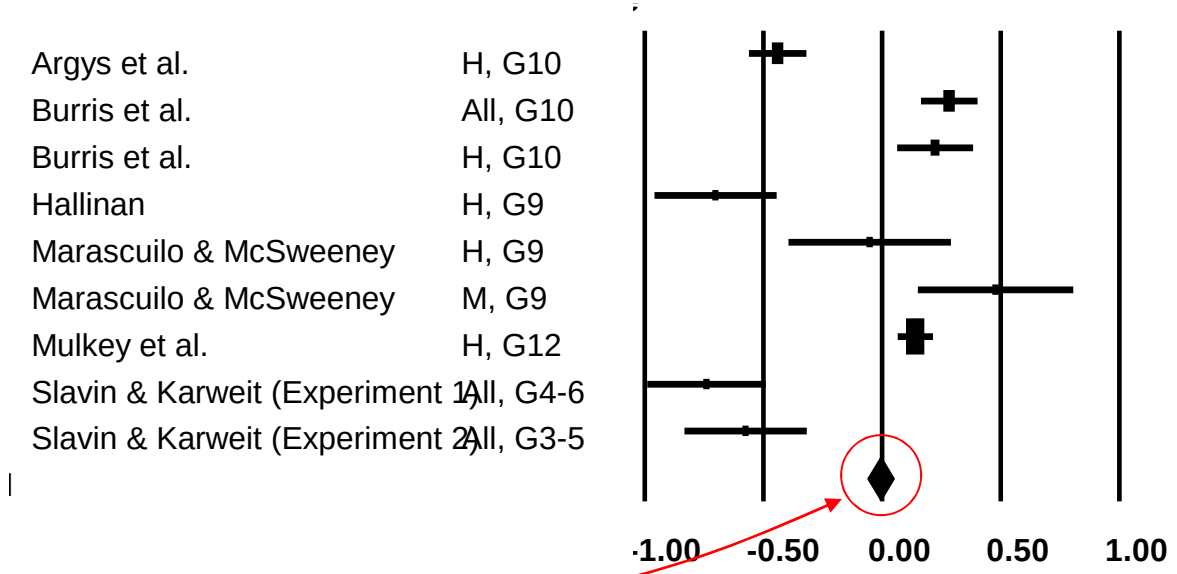
- Positive effects of detracking on student achievement for the 8 low-ability subgroups in 6 studies.
- Heterogeneity is reduced I^2 *(8)* > 82% without the outlier study
- Revealed by both fixed-effects model [***d* = .11**, *k*= 8, *p* <.0001] and random-effects model [***d* = .28**, *p* <.005]
- When only RCTs included, the impact is substantially higher, as revealed by a fixed-effects model (*d* = **0.63**, *k* = 4, *p* < 0.0001) and a random-effects model (*d* = **0.64**, *p* < 0.005).





Sensitivity Analysis

- One study (Kissoon-Singh, 1996), focused on average- and high-ability groups, reported extremely high ES (1.77 for high and 3.54 for average).
- Excluding the Kissoon-Singh (1996) study. Results indicate no effects of detracking in either direction for high- and average-ability students [***d* = -.005**, *p* =.837, 95% CI (-.053, .043)].



Discussion & Takeaways

- Meta-analysis extends previous research on the topic and offers new insights into this topic by synthesizing the best available evidence about the effects of a program or policy or relationship between two constructs from the literature.
- Meta-analysis is not panacea for biases from different studies, but is a good attempt to reduce those biases.
- The literature search and analysis procedure should as objective and clearly described as possible so that the results are replicable.
- Mata-analysis should strive for accuracy, clarity and simplicity.
- Downside: it can be time-consuming and labor-intensive.

Thank You!

Ning Rui

NingRui@Westat.com

Kathryn A. Henderson
KathrynHenderson2@westat.com