

Basic Aspects of Meta-Analysis

Hannah R. Rothstein

Baruch College- City University of New
York

Michael Borenstein

Biostat

Agenda

- Primary studies, systematic reviews and meta-analysis
- Basic components of a meta-analysis
- Effect Sizes
- Basic Calculations
- Heterogeneity of effects
- Comparison of fixed effect and random effects models

Part I

PRIMARY STUDIES, SYSTEMATIC REVIEWS AND META-ANALYSIS

Primary study

- Hypothesis
- Inclusion/exclusion criteria
- Collect data
- Statistical analysis (meta-analysis)
- Report

Systematic review

- Hypothesis
- Inclusion/exclusion criteria
- Collect data (information retrieval and data extraction)
- Statistical analysis (meta-analysis)
- Report

Meta-analysis

- Meta-analysis is the quantitative data analysis component of a systematic review.
- Not all systematic reviews include a meta-analysis.
- Not all meta-analyses are preceded by the earlier components of a systematic review.
- In practice, researchers will often use “systematic review” , “meta-analysis” and “research synthesis” as synonyms.

A Systematic Review Aims To Be

- Explicit (e.g. in its statement of objectives, materials and methods)
- Systematic (e.g. in its identification of literature)
- Transparent (e.g. in its criteria and decisions)
- Reproducible (e.g. in its methodology and conclusions)
- Unbiased

The move to systematic reviews and away from narrative reviews

Narrative reviews are not scientifically rigorous

- They use informal and subjective methods to collect and interpret information
- They generally provide a narrative summary of the research literature
- Different experts may perform a review on the same question and come to different conclusions
 - Sometimes due to review of different sets of studies
 - But can happen even when the same studies are reviewed

Narrative reviews become less efficient with more data

- Researcher may be able to combine the results of a few studies in his or her head
- Becomes increasingly difficult to do so as the number of studies increase
- We use statistics to combine information within a single study.....

Problems when the treatment effect varies

- As the number of studies grow, they often examine different populations
- The size of the relationship of interest may vary in different populations
- The narrative reviewer, who has enough trouble summarizing studies when they are all done in similar situations now has a much harder task

Too much literature

- Among the earliest meta-analyses were synthesis of:
 - 345 studies of the effects of interpersonal expectations on behavior (Rosenthal & Rubin, 1978),
 - 725 estimates of the relation between class size and academic achievement (Glass & Smith, 1979),
 - 833 tests of the effectiveness of psychotherapy (Smith & Glass, 1977),
 - 866 comparisons of the differential validity of employment tests for Black and White workers (Hunter, Schmidt & Hunter, 1979)

Is class size related to student achievement?

- By 1978 there were hundreds of studies on this topic

Narrative review

- Thompson concluded that the relationship ...involved too many complex issues to be reduced to a single testable hypothesis, and that research findings were “necessarily inconclusive”.

Meta-analysis

- Smith and Glass (1978)
 - 80 studies
 - 700 comparisons of smaller and larger classes
- Results showed clearly that smaller classes are better on
 - student achievement,
 - classroom processes
 - teacher and student attitudes.

Systematic vs. Narrative Reviews

- Scientific approach (models itself on primary research)
- Criteria determined a priori
- Comprehensive search for relevant information
- Explicit methods of data extraction and coding
- Meta-analysis generally be used to combine data
- Replicable
- Influenced by authors' point of view (bias)
- Author does not need to justify criteria for inclusion
- Search for data does not need to be comprehensive
- Methods not usually specified
- Narrative summary or vote count
- Can't replicate review

An attempted “fix”: Vote counting

- Collect a set of studies
 - Examine the tests of statistical significance
 - Tally the proportion significant as Yes votes
 - Tally the proportion not-significant as No votes
 - Majority wins
 - (There are variants with three categories: positive, negative and non-significant)

Warr and Perry (1982) Psych. Bulletin

Table 1
Summary of 38 Studies (Irrespective of Quality) Comparing Women's Psychological Well-Being and Their Paid Employment Status

Categories of women	Indices of psychological well-being											All indices	
	Suicide and attempted suicide (A)		Diagnosed psychiatric illness (B)		Psychiatric morbidity (C)		Psychological distress (D)		Life satisfaction or happiness (E)		Positive well-being (F)		
	+	ns	+	ns	+	ns	+	ns	+	ns	+		ns
Women in general (Groups 1, 2, 3, and 4)	2		2	1	2	3	2	4		1			8
Single women with no children at home (Group 1)					2		1						3
Single women in general (Groups 1 and 3)	1				1		1						3
Married women with no children at home (Group 2)							1	3					1
Married women with children at home (Group 4)							1	6		3		1	1
Other groups of women with children								3	1				1
Married women in general (Groups 2 and 4)	1			2		2	1	8		1			2
All of the above comparisons	4		2	3	5	5	7	24	1	5		1	19

Note. + = positive, ns = not significant. Comparisons identified as positive are those in which employed women have significantly higher psychological well-being than those who are unemployed. No cases of a negative association between employment status and psychological well-being were located.

Vote counting

- Focuses on the statistical significance of the primary studies
- Vote counting treats a nonsignificant p-value as evidence that an effect is absent. In fact, though, small, moderate, and even large effect sizes may yield a nonsignificant p-value due to inadequate statistical power.
- Often wrong, due to low power of primary studies

Vote-Counting --incorrect conclusions

- Hedges and Olkin (1980) showed that the power of vote-counting can not only be lower than that of the studies on which it is based, but can tend toward zero as the number of studies increases.
- With 20 studies of $N=30$ and an effect of $d= .5$ a vote count will fail to detect the effect 75% of the time
- In other words, vote counting is not only misleading, it tends to be more misleading as the amount of evidence increases!

Vote Counting

- Even when correct, doesn't provide information about the size of effects or the consistency of effects across studies
- Doesn't give more weight to more precise studies

Special Situations: Combined significance test and sign test

OTHER APPROACHES TO SYNTHESIS

Combined significance test meta-analysis (Rosenthal's early attempt)

- Advantage: lies in the increased power of the overall comparison.
- If several tests consistently favor the research question but fail to reach the level of significance, due to small sample sizes, the overall test is more likely to reach significance because the pooled sample size is much larger.
 - The hypothesis being tested is that the null hypothesis is true in every study. If this hypothesis is rejected, we can conclude that there is at least one study with a non null effect.
- It tells you nothing about the magnitude of effect size(s)

JCCP 1987 Shoham-Salomon & Rosenthal, Paradoxical Interventions

Table 3
Effect Sizes and Significance Levels of Comparisons Between Groups

Data set	Paradox type	N	Paradox vs. control		
			r	Z ^a	n
1	A	10	—	—	—
2	A	24	—	—	—
3	A	25	+.68	3.46	25
4	A	50	+.45	2.34	30
5	D	50	+.24	1.20	30
6	D	32	+.07	.31	22
7	D	30	+.46	2.40	30
8	D	43	+.27	1.66	43
9	C	30	+.72	3.58	20
10	C	29	+.54	2.46	20
11	B	29	+.40	1.75	20
12	E	29	+.20	.84	20
Unweighted \bar{r}			+.42		
Weighted \bar{r}			+.41		
Overall Z ^b				6.32	
Combined p				.0000001	

Note. See Table 1 for a description of paradox types. Signs designate higher (+) or lower (—) than control group (or two comparison groups and other for the third). Chi-square analysis was used to compare paradox vs. control group, $\chi^2(9) = 13.31$, $p = .16$; for the paradox vs. other group, $\chi^2(8) = 13.38$.

^a $Z = [df / \log_e(1 + t^2/df)]^{1/2} [1 - 1/2df]^{1/2}$ when $t = r \sqrt{1 - r^2} \times \sqrt{df}$ and df is based on:

^b Overall $Z = \sum Z_j / \sqrt{k}$.

Combining p Values

- Collect a set of studies
- Extract the p values from all of the tests of significance (whether significant or not)
- Compute an overall p value
- You obtain the overall strength of evidence that ***an effect*** exists
- Important Note: This test does not tell us anything about the value of an overall effect, or its statistical significance-
- This was popular in the early days of meta-analysis; Rosenthal recommended it in 1978.
- We have better methods now, but sometimes this is the best we can do.

Combining p values

- Advantages
 - You can use this if all you have is the results of significance tests
 - You can combine p values from any test statistic representing the substantive hypothesis of interest, even if the studies vary in design or analysis
 - A p value is a p value is a p value

Combining p values

- Disadvantages
 - Same issues as in primary studies
 - May have effect that is large but not statistically significant
 - particularly in primary studies
 - May have effect that is trivial, yet statistically significant
 - a particular problem for overall meta-analysis

A last resort: The sign test

- The sign test is used to count the number of studies with findings in one direction compared to the number of findings in the other direction, without regard to whether the findings are statistically significant
- If a treatment is completely ineffective, we expect that half of the studies will fall on each side of the no-effect line

The sign test

- Tests the hypothesis that the effect sizes from a collection of K independent studies are all zero
- Simply test if proportion of positive results is 50%
 - If the treatment has an effect, the probability of getting a positive result is greater than .5
 - If it has no effect, the probability of getting a positive result is .5

Sign test advantages

- The sign test is useful when
 - No numeric data are provided from studies, but directions of effects are provided
 - Numeric data are SO different that they cannot be combined statistically
 - Studies are so diverse in their populations or other characteristics that a pooled effect size is meaningless, but studies are addressing a questions sufficiently similar that the direction of effect is meaningful.
- Results can be tested for statistical significance using the standard binomial test

Sign test disadvantages

- Doesn't incorporate sample size (give more weight to more precise studies)
- Does not provide an estimate of effect size

Sign test: Example

- Health-related quality of life after liver transplantation: a meta-analysis (Bravata et al., 1999, *Liver Transp Surg*).
- “Performed a sign test on 49 studies to evaluate the direction (positive or negative) of the effect of transplantation on QOL. ”
- “The sign test showed significant improvement in posttransplantation physical health ($P < .0004$), sexual functioning ($P < .008$), daily activities ($P < .02$), general HRQL ($P < .02$), and social functioning ($P < .05$), but not psychological health ($P < .08$). ”

COMPONENTS OF A META-ANALYSIS

CLINICAL RESEARCH

Coronary Artery Disease

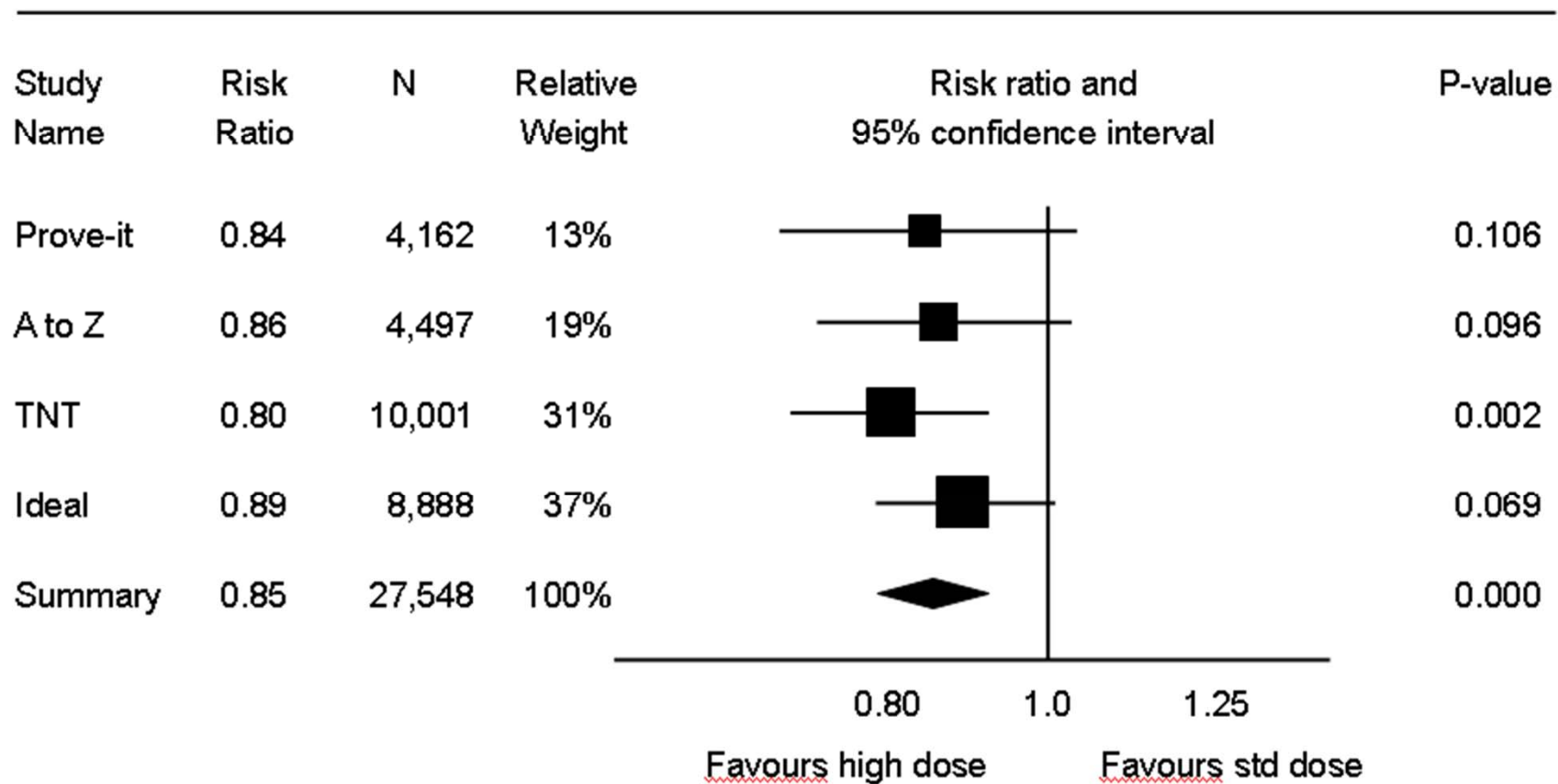
Meta-Analysis of Cardiovascular Outcomes Trials Comparing Intensive Versus Moderate Statin Therapy

Christopher P. Cannon, MD, Benjamin A. Steinberg, BA, Sabina A. Murphy, MPH,
Jessica L. Mega, MD, Eugene Braunwald, MD

Boston, Massachusetts

-
- OBJECTIVES** The purpose of this study was to conduct a meta-analysis that compares the reduction of cardiovascular outcomes with high-dose statin therapy versus standard dosing.
- BACKGROUND** Debate exists regarding the merit of more intensive lipid lowering with high-dose statin therapy as compared with standard-dose therapy.
- METHODS** We searched PubMed and article references for randomized controlled trials of intensive versus standard-dose statin therapy enrolling more than 1,000 patients with either stable coronary heart disease or acute coronary syndromes. Four trials were identified: the TNT (Treating to New Targets) and the IDEAL (Incremental Decrease in End Points Through Aggressive Lipid-Lowering) trials involved patients with stable cardiovascular disease, and the PROVE IT-TIMI-22 (Pravastatin or Atorvastatin Evaluation and Infection Therapy-Thrombolysis in Myocardial Infarction-22) and A-to-Z (Aggrastat-to-Zocor) trials involved patients with acute coronary syndromes. We carried out a meta-analysis of the relative odds on the basis of a fixed-effects model using the Mantel-Haenszel method for the major

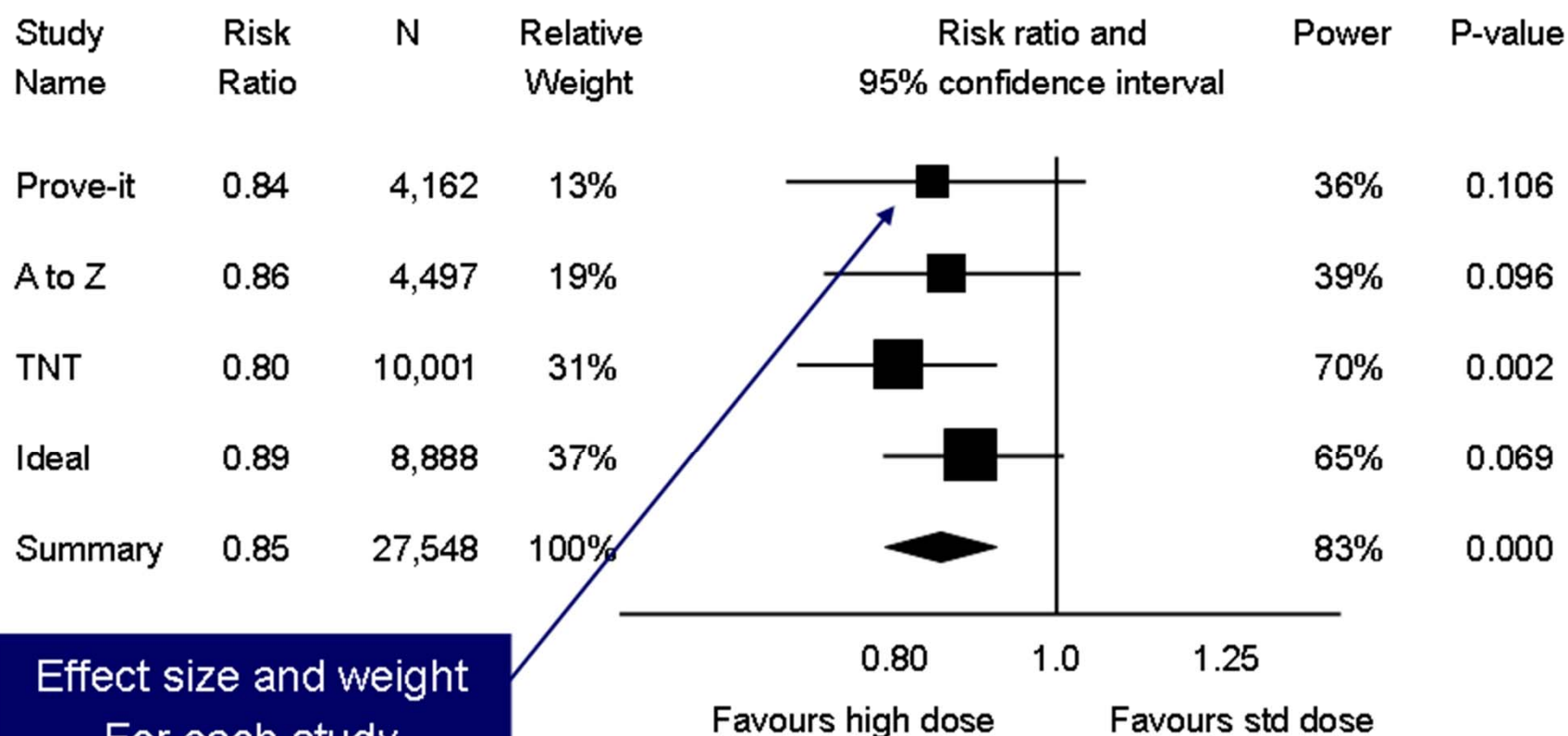
Impact of Statin Dose On Death and Myocardial Infarction



Effect size and weight

- Effect size can be based on means, proportions, and so on
- Weights based on amount of information carried by each study

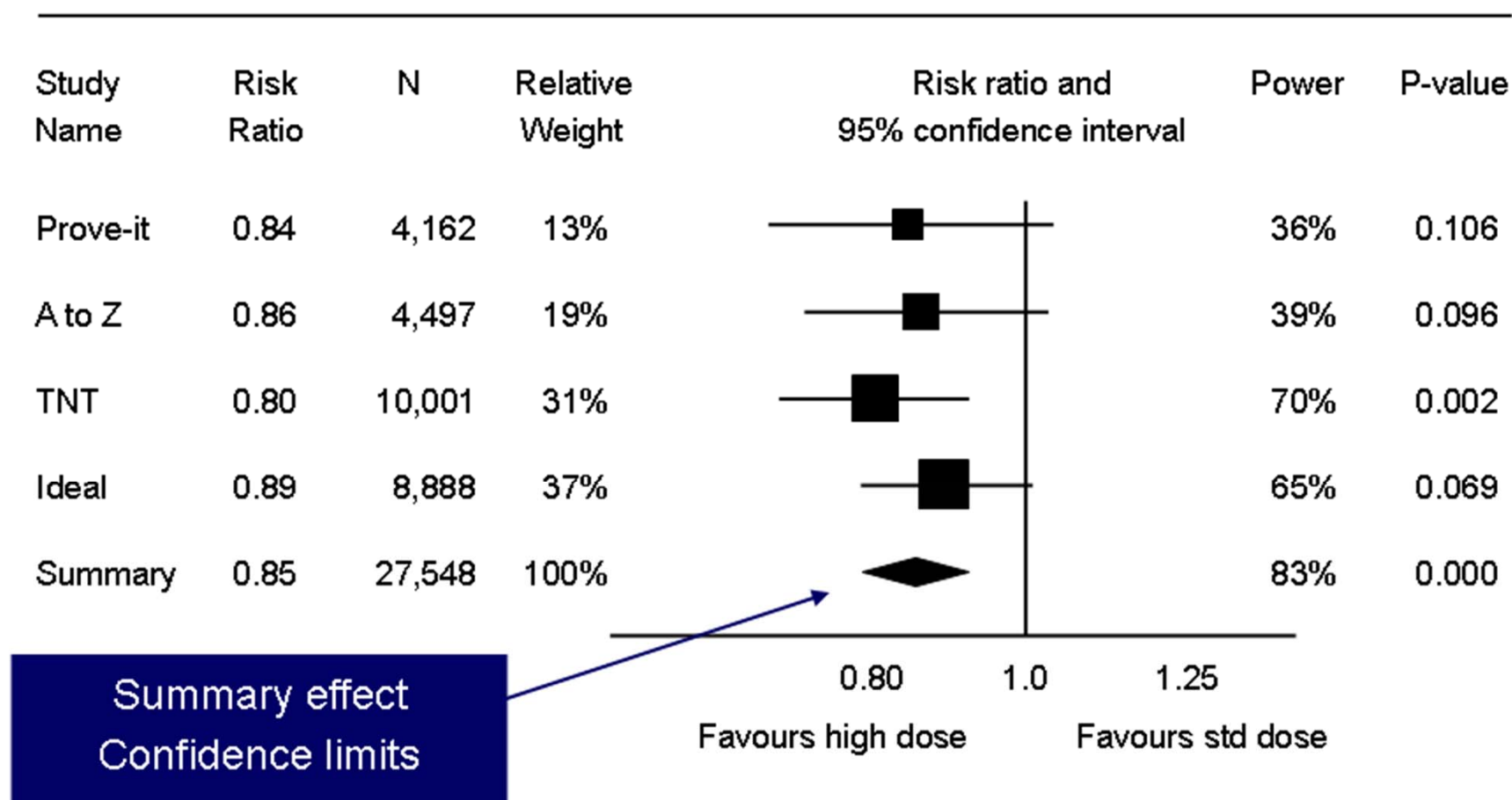
Impact of Statin Dose On Death and Myocardial Infarction



Summary effect and CI

- Summary effect is weighted mean
- Confidence interval based on same statistics as primary study
- Takes account of multiple levels of sampling

Impact of Statin Dose On Death and Myocardial Infarction



A meta-analysis is a synthesis

- How we approach the synthesis depends on our research question
- What kinds of studies we include depends on our research question
- The kind of analysis depends on our research question
- The conclusions depend on our research question and the information available

Compared to primary study

- Provides a context
 - The treatment effect is similar across the set of studies
 - The treatment effect varies in unknown ways
 - The treatment effect varies in ways we can model

When the treatment effect is consistent across studies

Meta-analysis

- Provides a more precise estimate of effect size
- Provides increased statistical power

Streptokinase

- A now classic systematic review and meta-analysis

Model	Study name	Year	Statistics for each study					Odds ratio and 95% CI				
			Odds ratio	Lower limit	Upper limit	Z-Value	p-Value	0.01	0.10	1.00	10.00	100.00
	Dewar	1963	0.471	0.114	1.942	-1.042	0.297					
	European 1	1969	1.460	0.689	3.096	0.987	0.323					
	European 2	1971	0.635	0.447	0.903	-2.529	0.011					
	Heikinheimo	1971	1.248	0.643	2.423	0.655	0.513					
	Italian	1971	1.012	0.510	2.008	0.034	0.973					
	Australian 1	1973	0.754	0.436	1.306	-1.006	0.314					
	Franfurt 2	1973	0.378	0.183	0.778	-2.640	0.008					
	NHLBI	1974	2.587	0.632	10.596	1.321	0.186					
	Frank	1975	0.959	0.289	3.185	-0.068	0.946					
	Valere	1975	1.061	0.392	2.876	0.117	0.907					
	Klein	1976	3.200	0.296	34.588	0.958	0.338					
	UK-Collab	1976	0.910	0.565	1.466	-0.386	0.699					
	Austrian	1977	0.562	0.365	0.867	-2.609	0.009					
	Australian 2	1977	0.625	0.341	1.147	-1.518	0.129					
	Laserra	1977	0.222	0.019	2.533	-1.211	0.226					
	N Ger	1977	1.215	0.797	1.853	0.906	0.365					
	Witchitz	1977	0.778	0.199	3.044	-0.361	0.718					
	European 3	1979	0.561	0.298	1.055	-1.794	0.073					
	ISAM	1986	0.872	0.599	1.270	-0.713	0.476					
	GISSI-1	1986	0.807	0.721	0.903	-3.741	0.000					
	Olson	1986	0.407	0.035	4.795	-0.714	0.475					
	Baroffio	1986	0.064	0.003	1.192	-1.843	0.065					
	Schreiber	1986	0.296	0.028	3.142	-1.010	0.313					
	Cribier	1986	1.100	0.064	18.774	0.066	0.948					
	Sainsous	1986	0.467	0.110	1.986	-1.030	0.303					
	Durand	1987	0.586	0.120	2.861	-0.661	0.509					
	White	1987	0.159	0.035	0.727	-2.371	0.018					
	Bassand	1987	0.571	0.157	2.080	-0.849	0.396					
	Vlay	1988	0.417	0.033	5.299	-0.675	0.500					
	Kennedy	1988	0.631	0.292	1.362	-1.174	0.241					
	ISIS-2	1988	0.746	0.676	0.822	-5.877	0.000					
	Wisenberg	1988	0.205	0.037	1.153	-1.799	0.072					
Fixed			0.768	0.720	0.819	-8.007	0.000					

Fixed Random Both models

Basic stats One study removed Cumulative analysis Calculations

Basic Analysis

- Treatment effect was consistent
- Combined effect was substantial
- Only 6 of 33 studies met criterion for significance

Comprehensive meta analysis - [Analysis]

File Edit Format View Computational options Analyses Help

← Data entry ↗ Next table ⚙ High resolution plot 📄 Select by ... + Effect measure: Odds ratio

📊 📈 📉 📊 📉 📊 📉 📊 📉

Model	Study name	Cumulative statistics			Year	Cumulative Events / Total		Cumulative odds ratio (95% CI)		
		Point	Z-Value	p-Value		Group-A	Group-B	0.50	1.00	2.00
	Dewar	0.355	-1.667	0.096	1963	5 / 33	11 / 32			
	European 1	0.989	-0.034	0.973	1969	25 / 116	26 / 116			
	European 2	0.704	-2.233	0.026	1971	94 / 489	120 / 473			
	Italian	0.748	-2.022	0.043	1971	113 / 653	138 / 630			
	Heikinheimo	0.809	-1.607	0.108	1971	135 / 872	155 / 837			
	Franfurt 2	0.742	-2.403	0.016	1973	148 / 974	184 / 941			
	Australian 1	0.744	-2.604	0.009	1973	174 / 1238	216 / 1194			
	NHLBI	0.767	-2.366	0.018	1974	181 / 1291	219 / 1248			
	Frank	0.772	-2.340	0.019	1975	187 / 1346	225 / 1301			
	Valere	0.783	-2.262	0.024	1975	198 / 1395	234 / 1343			
	UK-Collab	0.803	-2.224	0.026	1976	236 / 1697	274 / 1636			
	Klein	0.810	-2.139	0.032	1976	240 / 1711	275 / 1645			
	Lasiera	0.804	-2.228	0.026	1977	241 / 1724	278 / 1656			
	Austrian	0.758	-3.094	0.002	1977	276 / 2076	343 / 2032			
	Australian 2	0.747	-3.394	0.001	1977	303 / 2199	374 / 2139			
	Witchitz	0.747	-3.413	0.001	1977	308 / 2231	379 / 2165			
	N Ger	0.798	-2.838	0.005	1977	371 / 2480	430 / 2399			
	European 3	0.782	-3.185	0.001	1979	389 / 2636	460 / 2558			
	Baroffio	0.777	-3.276	0.001	1986	389 / 2665	466 / 2588			
	Schreiber	0.774	-3.333	0.001	1986	390 / 2684	469 / 2607			
	Olson	0.772	-3.371	0.001	1986	391 / 2712	471 / 2631			
	Sainsous	0.768	-3.459	0.001	1986	394 / 2761	477 / 2680			
	GISSI-1	0.792	-5.069	0.000	1986	1022 / 8621	1235 / 8532			
	ISAM	0.797	-5.096	0.000	1986	1076 / 9480	1298 / 9414			
	Cribier	0.797	-5.092	0.000	1986	1077 / 9501	1299 / 9437			
	White	0.793	-5.219	0.000	1987	1079 / 9608	1311 / 9549			
	Bassand	0.791	-5.265	0.000	1987	1083 / 9660	1318 / 9604			
	Durand	0.791	-5.293	0.000	1987	1086 / 9695	1322 / 9633			
	Wisenberg	0.788	-5.377	0.000	1988	1088 / 9736	1327 / 9658			
	Vlay	0.787	-5.397	0.000	1988	1089 / 9749	1329 / 9670			
	Kennedy	0.785	-5.494	0.000	1988	1101 / 9940	1346 / 9947			
	ISIS-2	0.768	-8.007	0.000	1988	1892 / 18532	2375 / 18442			
Fixed		0.768	-8.007	0.000						

Fixed Random

Split, Croatia June 2014

46

Basic stats One study removed Cumulative analysis Calculations

Cumulative Analysis

- Meta-analysis in 1977 could have been definitive
- 40,000 patients randomized after 1977
- Additional millions not treated
- Even in 1992, narrative review was not definitive

Studies vary in unknown ways

Meta-analysis

- Provides more precise estimate of effect size
- Provides increased statistical power
- Allows us to measure the heterogeneity of the effect

Posttraumatic Stress Disorder in Parents of Children With Chronic Illnesses: A Meta-Analysis

Mariana Cabizuca
Universidade Federal do Rio de Janeiro

Carla Marques-Portella
Universidade Federal do Rio de Janeiro

Mauro V. Mendlowicz
Universidade Federal Fluminense

Evandro S. F. Coutinho
Escola Nacional de Saúde Pública-Fundação Oswaldo Cruz

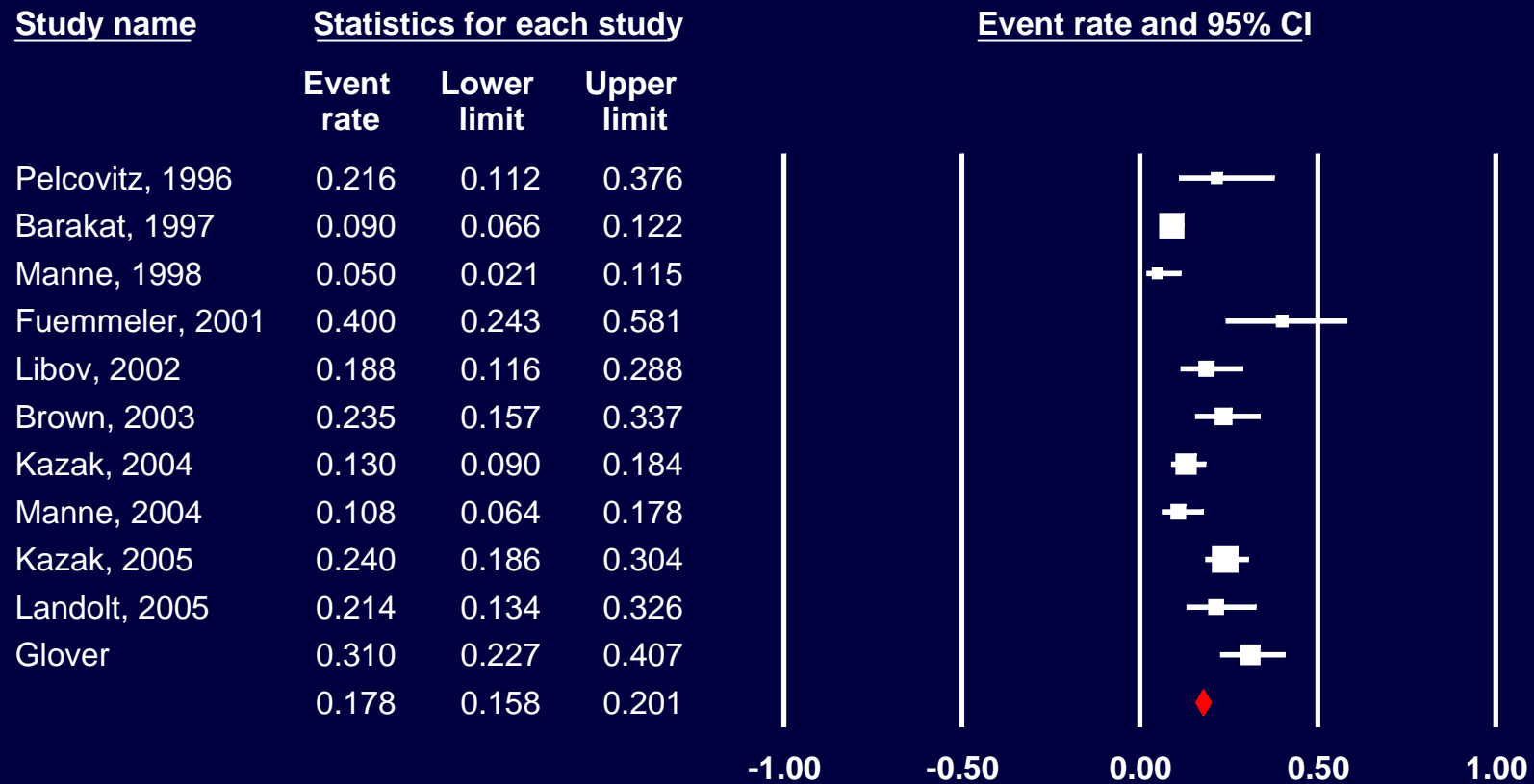
Ivan Figueira
Universidade Federal do Rio de Janeiro

Objective: To estimate PTSD prevalence in parents of children with chronic illnesses or undergoing invasive procedures, and its association with higher risk of PTSD among parents. *Methods:* Sixteen studies reporting prevalence of PTSD in parents of children with chronic illnesses were identified through a systematic review in Pubmed, Web of Science, Pilots and Psycinfo databases. *Main Outcome Measures:* Pooled current PTSD prevalence was calculated for parents from these studies. Pooled PTSD prevalence ratios were obtained by comparing parents of children with chronic diseases with parents of healthy children. Meta-regression was used to identify variables that could account for the lack of homogeneity. *Results:* Pooled PTSD prevalence was 19.6% in mothers, 11.6% in fathers, and 22.8% in

Prevalence

- Based on the heterogeneity index (tau-squared), most of the variation here is real; that is it is not due to chance.
- Prevalence varied from 5% to 40% across studies; the reasons for this heterogeneity were not investigated.

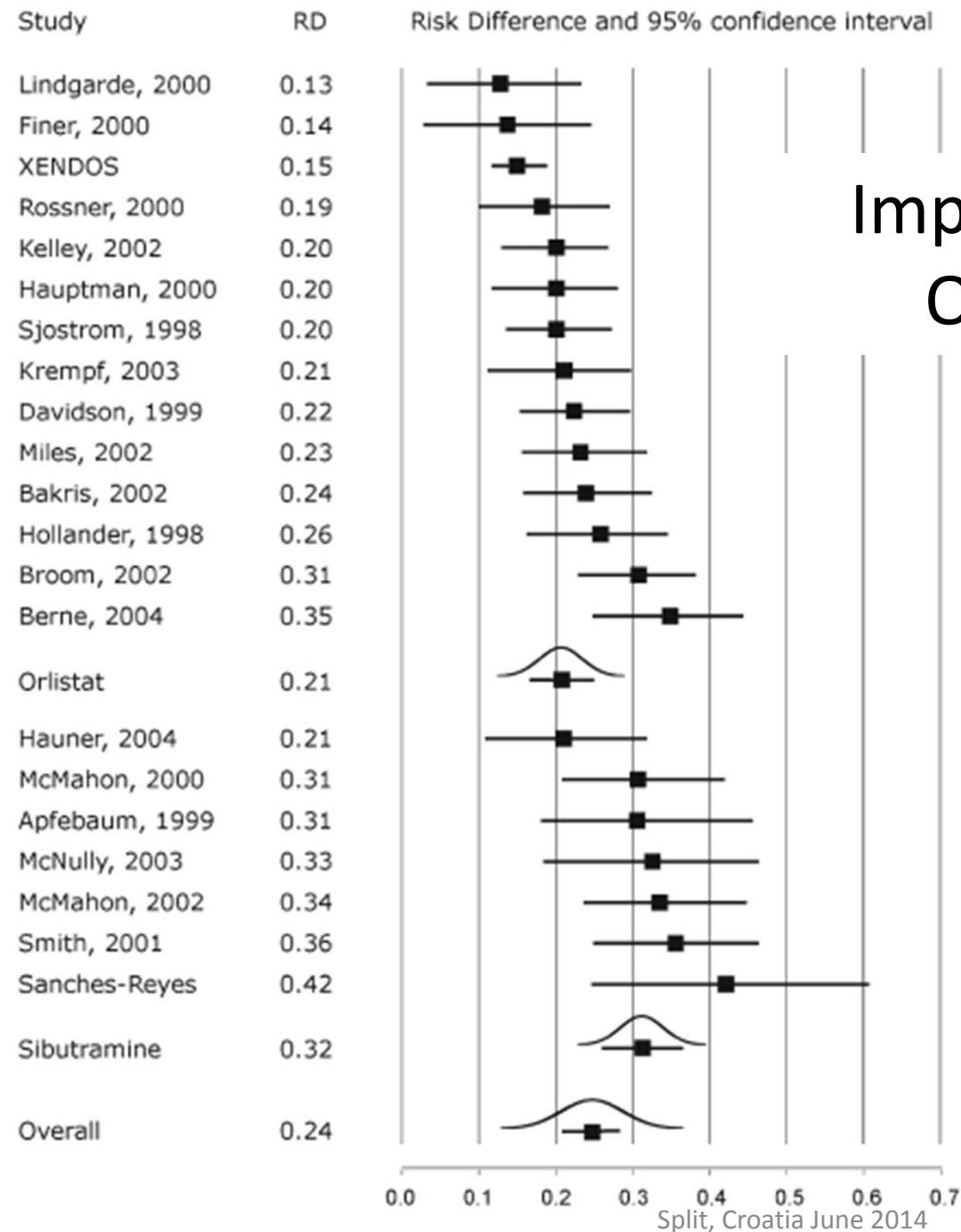
Pevalence of PTSD



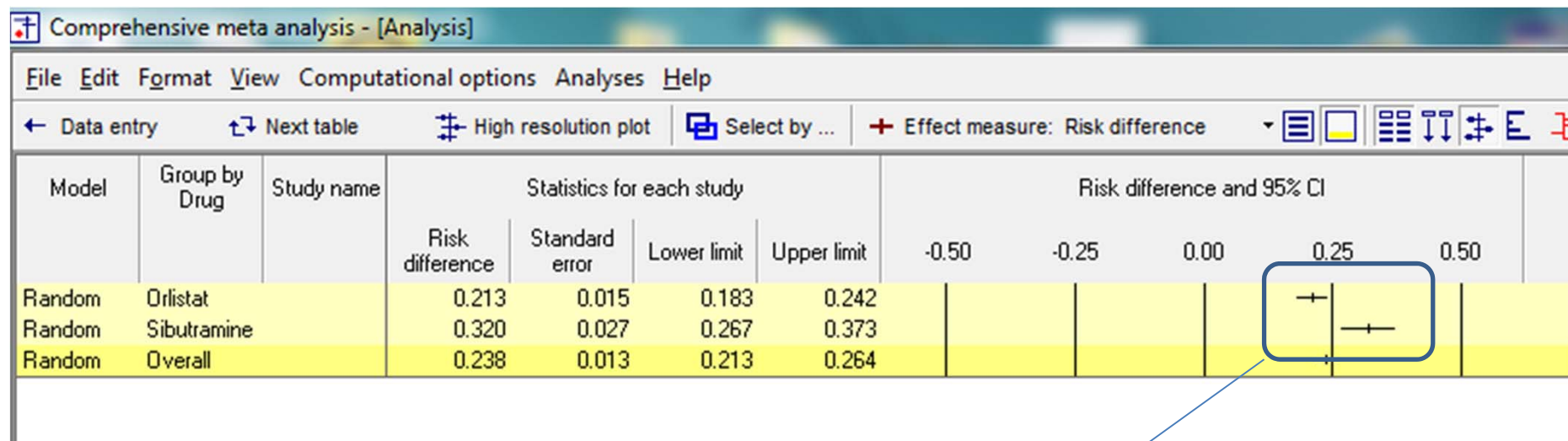
Studies vary in ways we can model

Meta-analysis

- Allows us to assess treatment effect (and variation) within subgroups
- Assess us to measure the impact of moderator(s) on the treatment effect



Impact of two Drugs On Weight Loss



Groups		Effect size and 95% confidence interval					Test of null (2-Tail)		Heterogeneity		
Group	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value
Fixed effect analysis											
Orlistat	14	0.200	0.010	0.000	0.180	0.219	20.236	0.000	27.560	13	0.010
Sibutramine	7	0.319	0.022	0.001	0.275	0.363	14.173	0.000	6.454	6	0.374
Total within									34.014	19	0.018
Total between									23.532	1	0.000
Overall	21	0.219	0.009	0.000	0.201	0.236	24.225	0.000	57.546	20	0.000
Mixed effects analysis											
Orlistat	14	0.213	0.015	0.000	0.183	0.242	14.102	0.000	12.098	1	0.001
Sibutramine	7	0.320	0.027	0.001	0.267	0.373	11.853	0.000			
Total between											
Overall	21	0.238	0.013	0.000	0.213	0.264	18.091	0.000			

Outcomes matter

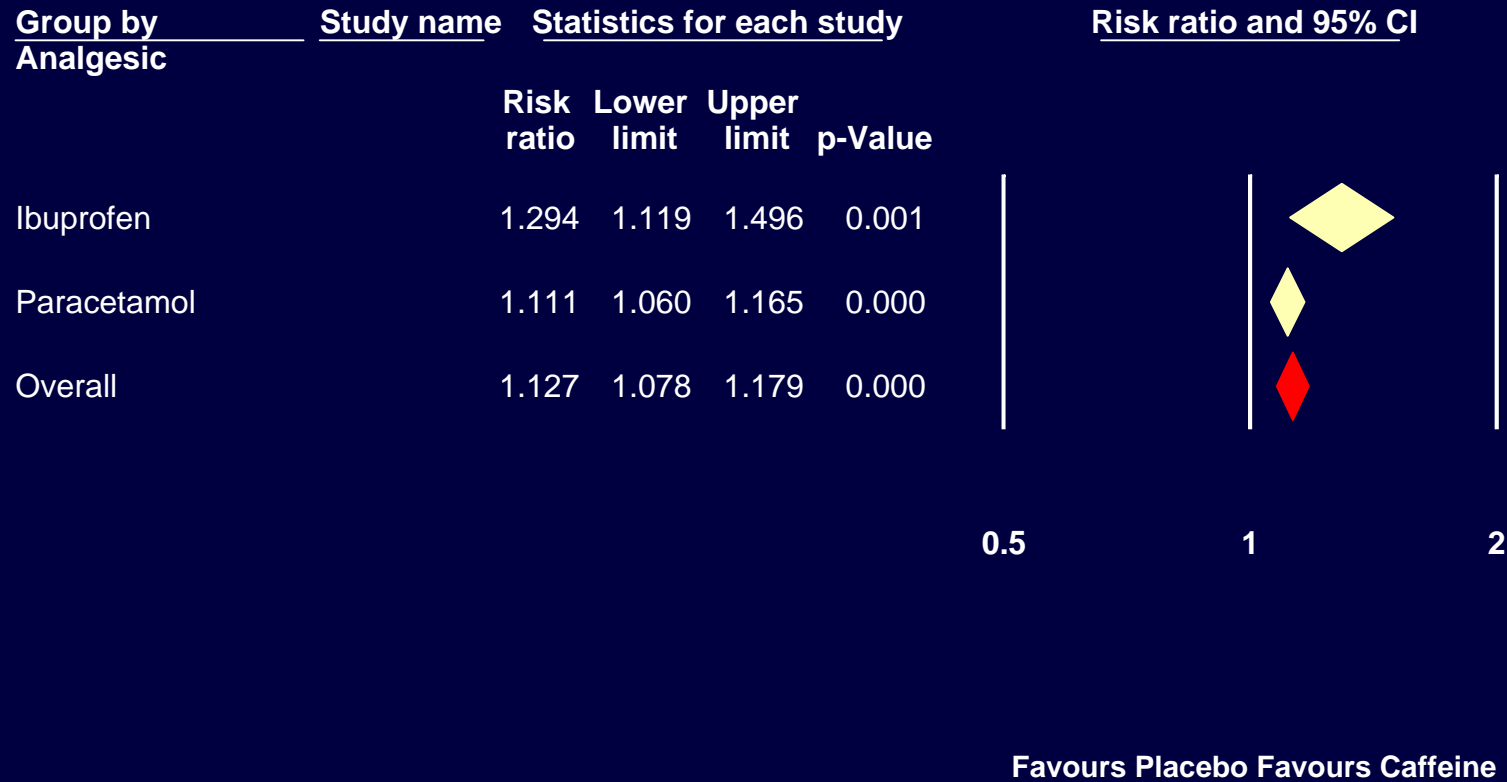
- These two obesity drugs have different mechanisms, different side effects and different contraindications.
- If we had looked at one of the side effects as an outcome, we might have a different picture.

Caffeine as an analgesic adjuvant for acute pain in adults (Review)

Derry CJ, Derry S, Moore RA



Caffeine by Analgesic | Relief



Meta Analysis

Meta-analysis of large randomized controlled trials to evaluate the impact of statins on cardiovascular outcomes

Bernard M. Y. Cheung, Ian J. Lauder,¹ Chu-Pak Lau & Cyrus R. Kumana

Department of Medicine and ¹Statistics and Actuarial Science, University of Hong Kong, Queen Mary Hospital, Hong Kong

Correspondence

Dr Bernard M. Y. Cheung, Associate Professor, University Department of Medicine, Queen Mary Hospital, Hong Kong.

Tel: + 852 2855 4768

Fax: + 852 2904 9443

Aims

Since 2002, there have been five major outcome trials of statins reporting findings from more than 47 000 subjects. As individual trial results differed, we performed a meta-analysis to ascertain the effectiveness and safety of statins overall and in subgroups. The aim of the study was to estimate the effect of statins on major coronary events and strokes, all-cause mortality and noncardiovascular mortality, and in different subgroups.

Impact of Statins by Smoking

Comprehensive meta analysis - [Analysis]

File Edit Format View Computational options Analyses Help

← Data entry ↔ Next table ⚙ High resolution plot 📄 Select by ... + Effect measure: Risk ratio

</

Bottom line

- Can limit the analysis to studies that are essentially identical, and get more precise estimate of the common effect
- Can include studies that vary in random ways, assess variation in effect
- Can include studies that vary on potentially important factors, assess the impact of these factors

Caution about moderator analysis

- Moderator variables use up power
- Too many moderator analyses lead to Type I errors
- Variables are often confounded with each other
- Moderator analyses are always observational—they don't support causal inferences

Effect-size indices

Indices

- Continuous Data
- Dichotomous (Binary) Data
- Correlations
- Others (E.g. prevalence, mean in one group)

Indices for means

- Raw mean difference D
- Standardized mean difference d and g

Raw mean difference

$$D = \overline{X}_1 - \overline{X}_2$$

$$D = 550 - 500 = 50$$

Standardized Mean Difference

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S_{Within}}$$

$$d = \frac{550 - 500}{100} = 0.50$$

Raw vs. Standardized Difference

	D	d (or g)
Scale is natural or known	Required	Not required
All studies must use same scale	Required	Not required
Standard deviation must be consistent	Required	Not required

Indices for binary outcomes

- Definition of **risk**
- Definition of **odds**
- Risk difference *RD*
- Risk ratio (relative risk) *RR*
- Odds ratio *OR*

Table 5.1 Nomenclature for 2×2 table of outcome by treatment.

	Events	Non-Events	N
Treated	A	B	n_1
Control	C	D	n_2

Table 5.2 Fictional data for a 2×2 table.

	Dead	Alive	N
Treated	5	95	100
Control	10	90	100

Meaning of risk

- The number of people with the event/condition we are interested in compared with the total number of people who could have had it.
- Example from Cochrane
 - 24 people drank Coffee; 6 developed a headache.
 - The **Risk** of developing a headache (given coffee drinking) is $6/24$ or 25%

Meaning of Odds

- The number of people **with** the event of interest compared with the number **without** the event of interest.
- In the running Cochrane example, if 24 people drank coffee and 6 developed a headache
- The Odds of developing a headache (given coffee drinking) are 6/18, or 1 in 3.

Risk versus Odds

IN GENERAL

- When an event is rare, there won't be too much difference between the risk and the odds
- When an event is common, there can be a BIG difference between the risk and the odds.

Risk Difference

- Compute the risk of the event in the treated group
- Compute the risk of the event in the control group
- The Risk difference is the Risk in the treated group MINUS the Risk in the control group
- A Risk difference of ZERO means that there is no difference between the groups

Risk Difference

$$RD = \left(\frac{A}{n_1} \right) - \left(\frac{C}{n_2} \right)$$

$$RD = \left(\frac{5}{100} \right) - \left(\frac{10}{100} \right) = -0.05$$

Risk Ratio

- Compute the risk in the treated group
- Compute the risk in the control group
- The Risk Ratio is the risk in the treated group divided by the risk in the control group
- When the risk ratio is 1, that means there is NO difference between the groups
- When the risk ratio is above 1, it means that the risk is higher in the treated group than in the control group.

Risk Ratio (Relative Risk)

$$RR = \frac{A / n_1}{C / n_2}$$

$$RR = \frac{5 / 100}{10 / 100} = 0.50$$

Risk ratio

$$\frac{5 / 100}{10 / 100} = 0.50$$

$$\frac{0.50 + 2.00}{2} = 1.25$$

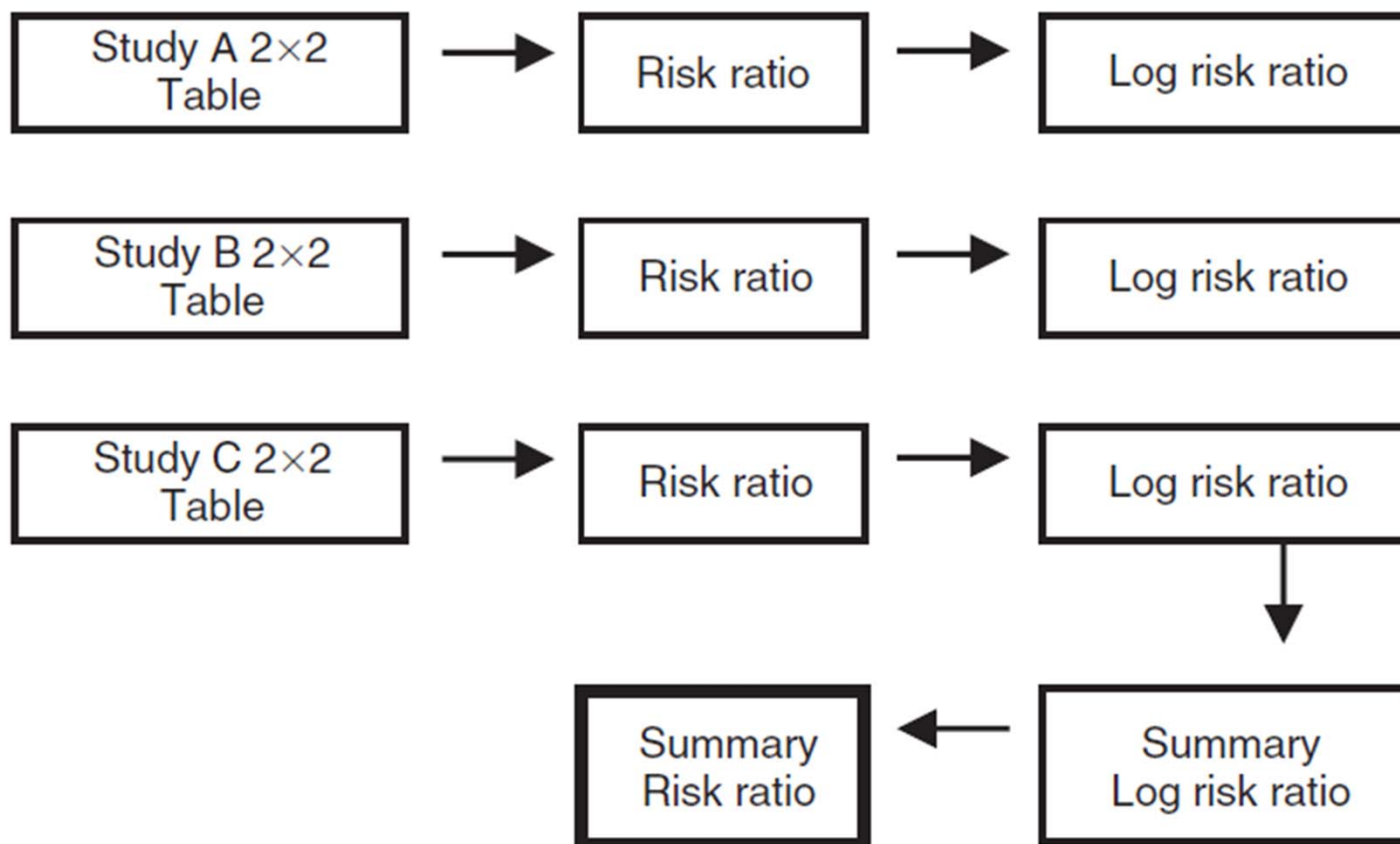
$$\frac{10 / 100}{5 / 100} = 2.00$$

Log Risk ratio

$$\ln(0.50) = -0.693$$

$$\frac{-0.693 + 0.693}{2} = 0.00$$

$$\ln(2.00) = +0.693$$

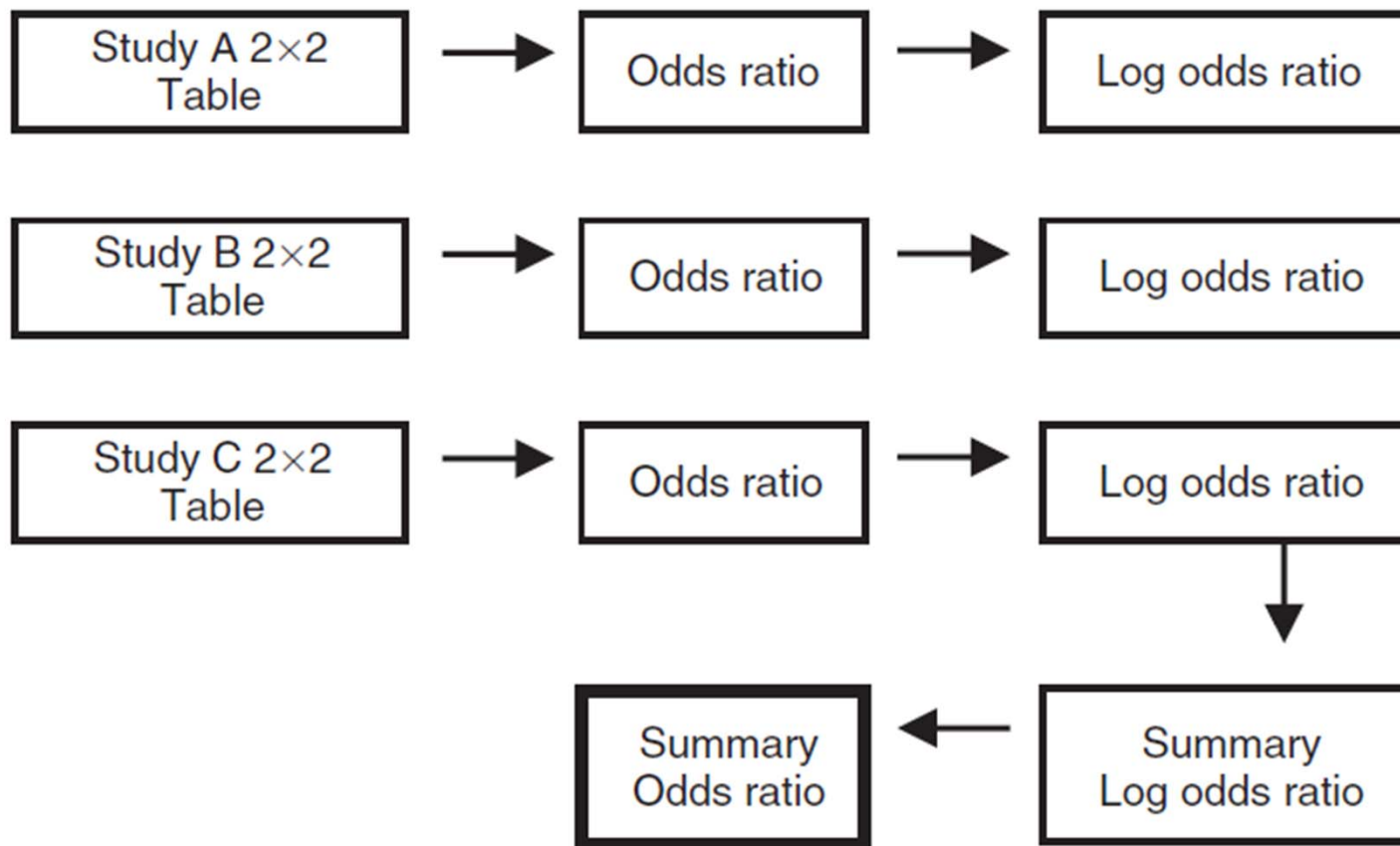


- Compute the odds in the treated group
- Compute the odds in the control group
- The Odds Ratio is the odds in the treated group divided by the odds in the control group
- When the odds ratio is 1, that means there is NO difference between the groups

Odds Ratio

$$OR = \frac{AD}{BC}$$

$$OR = \frac{5 \times 90}{95 \times 10} = 0.47$$



Research Ties Diabetes Drug to Heart Woes

By GARDINER HARRIS
Published: February 19, 2010

Hundreds of people taking [Avandia](#), a controversial [diabetes](#) medicine, needlessly suffer heart attacks and [heart failure](#) each month, according to confidential government reports that recommend the drug be removed from the market.


 [Enlarge This Image](#)




The reports, obtained by The New York Times, say that if every diabetic now taking Avandia were instead given a similar pill named Actos, about 500 heart attacks and 300 cases of heart failure would be averted every month because Avandia can hurt the heart. Avandia, intended to treat [Type 2 diabetes](#), is known as rosiglitazone and was linked to 304 deaths during the third quarter of 2009.

 TWITTER


 LINKEDIN

 COMMENTS
(162)

 E-MAIL

 PRINT

 REPRINTS

 SHARE

THE
WAY WAY BACK
WATCH TRAILER


Research Ties Diabetes Drug to Heart Woes

By GARDINER HARRIS

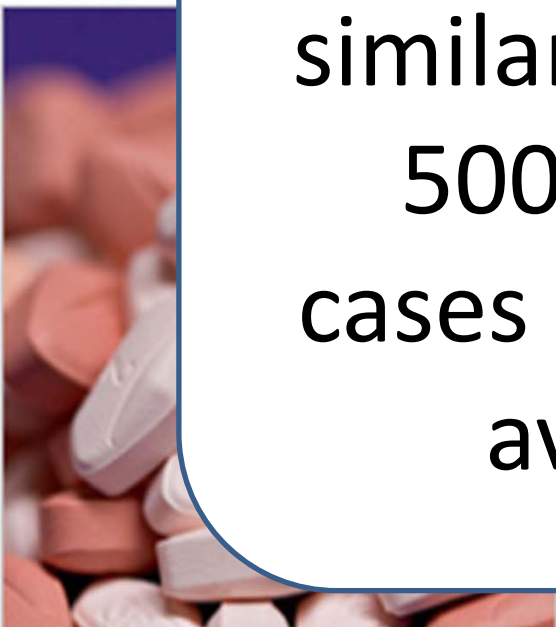
Published: February 19, 2010

Hundreds of people taking Avandia, a controversial diabetes medicine, a month, a recomme

 TWITTER

 LINKEDIN

“If every diabetic now taking Avandia were instead given a similar pill named Actos, about 500 heart attacks and 300 cases of heart failure would be averted every month”



F.D.A. to Restrict Avandia, Citing Heart Risk



Joe Raedle/Getty Images

A bottle of the controversial diabetes drug Avandia.

By GARDINER HARRIS


Published: September 23, 2010

WASHINGTON — In a highly unusual coordinated announcement, drug regulators in Europe and the United States said Thursday that Avandia, the controversial diabetes medicine, would no longer be widely available.

 RECOMMEND

 TWITTER

 LINKEDIN

 E-MAIL

Log
are
Priv

Wh

Dra
Bos
Sus
Cap



Life,
April 1

Rice,
April 1

Hea
April 1

Low
April 1

HPV
April 1

1
h

F.D.A. to Restrict Avandia, Citing Heart Risk

“In a highly unusual coordinated announcement, drug regulators in the United States and Europe said Thursday that Avandia ... would no longer be widely available

drug regulators in Europe and the United States said Thursday that Avandia, the controversial diabetes medicine, would no longer be widely available.

TWITTER

LINKEDIN

E-MAIL

Choice of effect size index: Avandia

- RR is 1.43 – your chances of dying increase by almost 50%
- RD is .005– the risk goes up from one in a thousand, versus 1 ½ per thousand.

RD vs. RR vs. OR

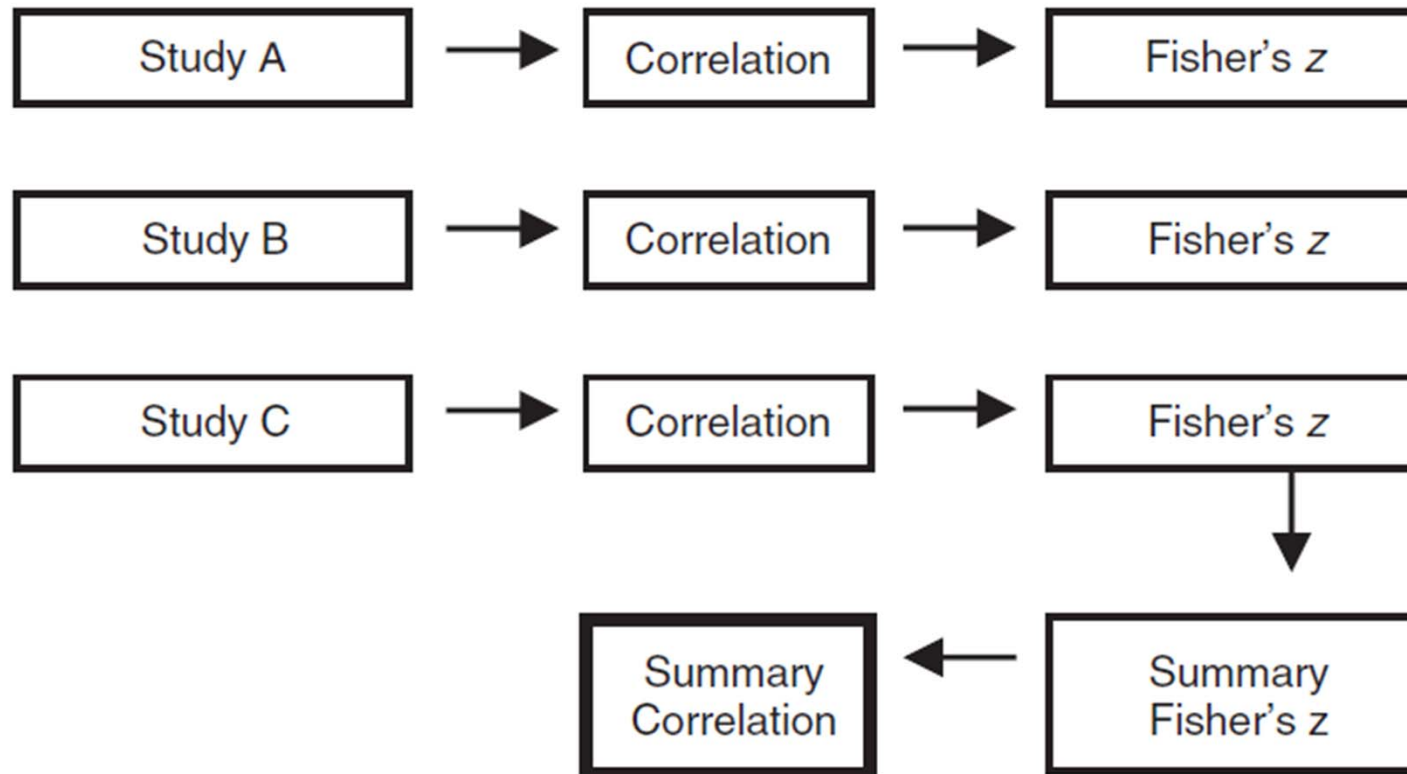
	<i>RD</i>	<i>RR</i>	<i>OR</i>
Type of difference	Absolute	Relative	Relative
Intuitive for	Researchers	Researchers	Statisticians
Special statistical issues	Prob. of failure and prob of success are reciprocal	Prob of failure and of success NOT reciprocal, and weights vary by choice	Prob. of failure and prob of success are reciprocal
Metric for analyses	Raw	Log	Log

Cochrane likes the Risk Difference

- It is easy to compute
- It is an **absolute** measure of actual change in risk
- It is easy to convert to natural frequencies and to NNT
- HOWEVER: The Risk Difference is more variable than the RR or OR across different populations, when baseline risk varies.

Correlation

- Used when both predictor and outcome variable are continuous.
- Used mostly in observational studies.



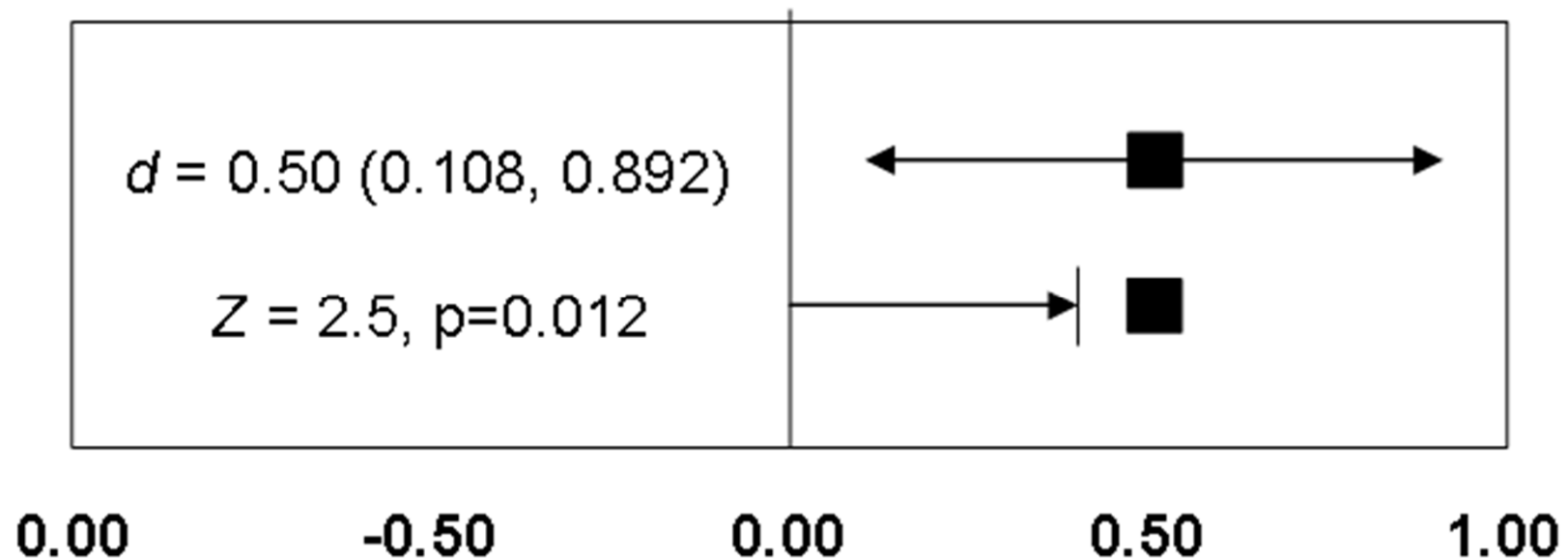
Effect Sizes Conventions

- Uses and cautions

Effect sizes
rather than p -values

p -value and effect size

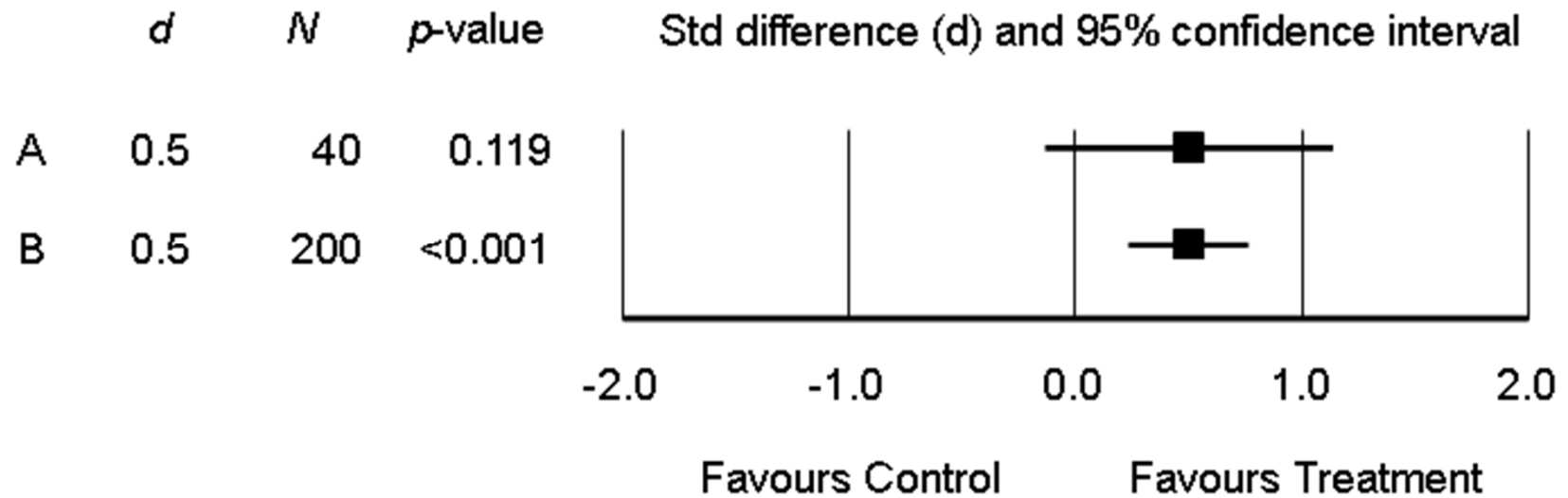
Approaches are consistent



Example-1

- One study $p = .119$
- One study $p = .001$
- Which study had the larger effect size?

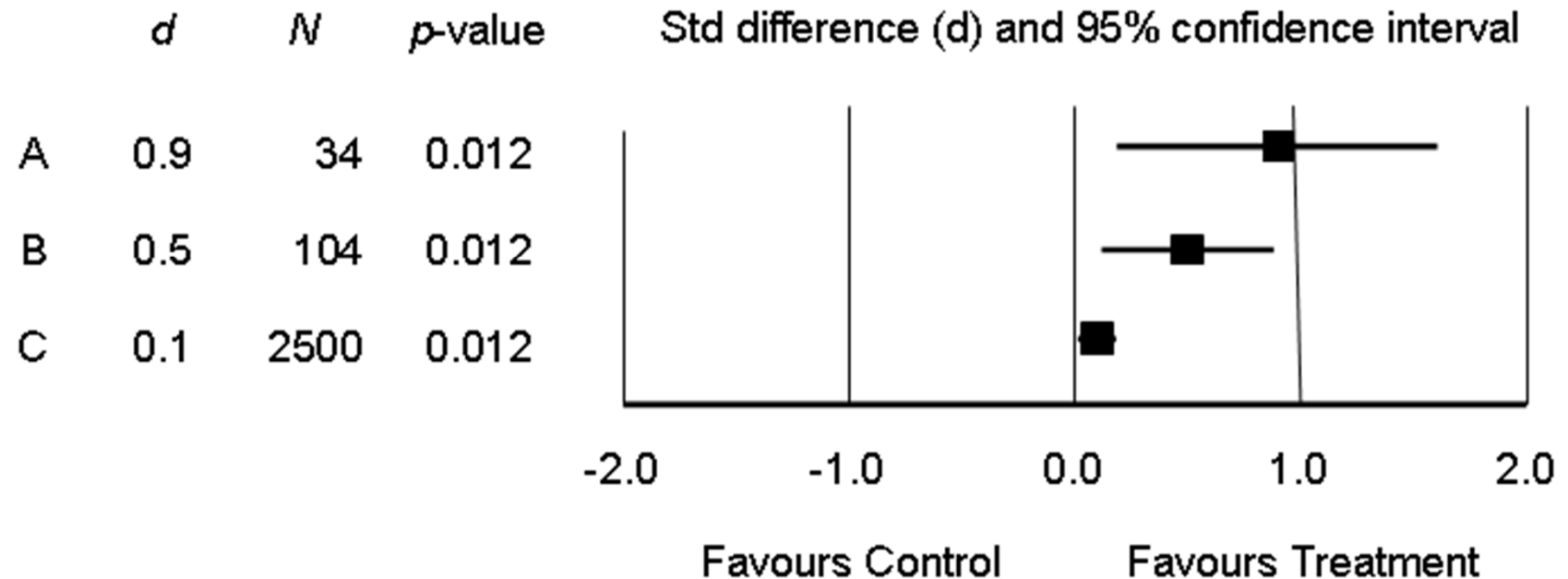
Example 1



Example 2

- One study $p = .012$
- One study $p = .012$
- One study $p = .012$
- Which study had the larger effect size?

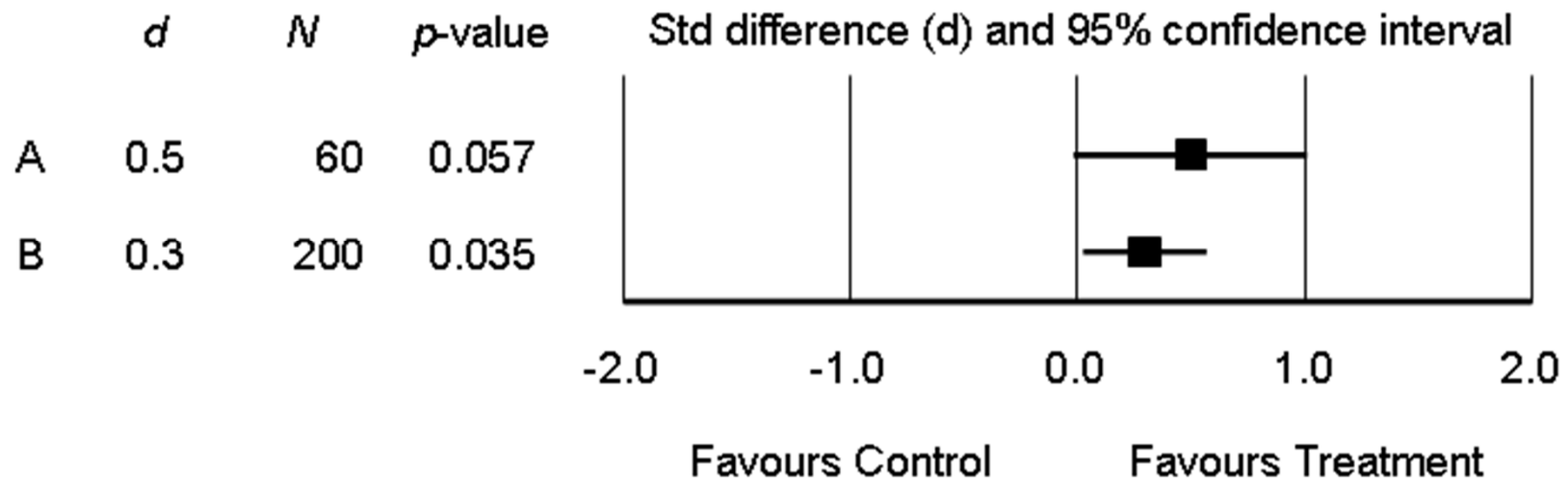
Example 2



Example 3

- One study $p = .057$
- One study $p = .035$
- Which study had the larger effect size?

Example 3



Key point

- Always important to work with ES
- In meta-analysis, especially so

Key point

- p -value is poor surrogate for effect size
- In primary studies, should report ES
- In meta-analyses must work with ES

Three studies, non significant

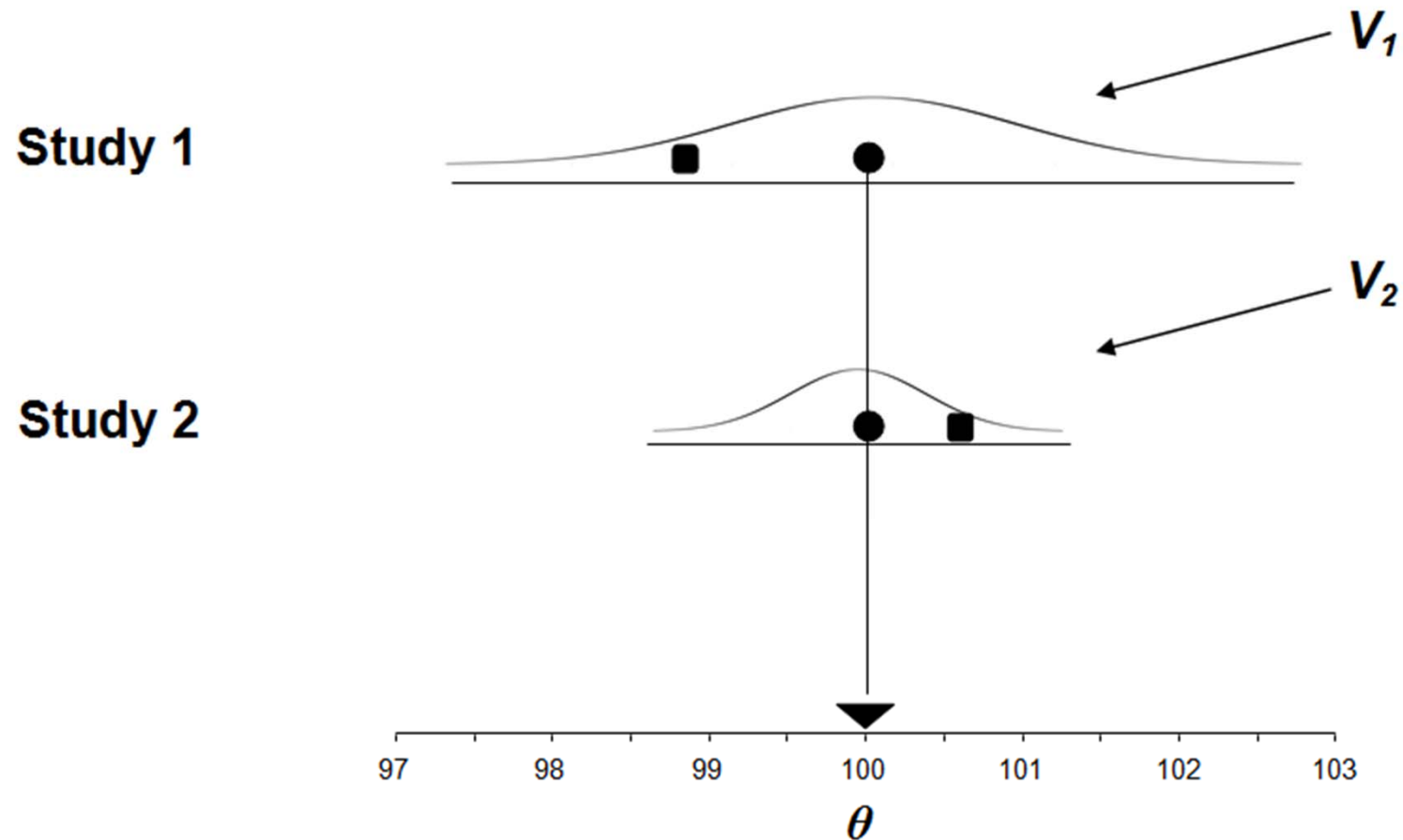
- Focus on p-values, synthesize conclusions and might conclude is no evidence of effect
- Focus on effect size, synthesize effects, as in next slide

Meta-Analysis
with consistent effects

FIXED-EFFECT META-ANALYSIS

Sampling error

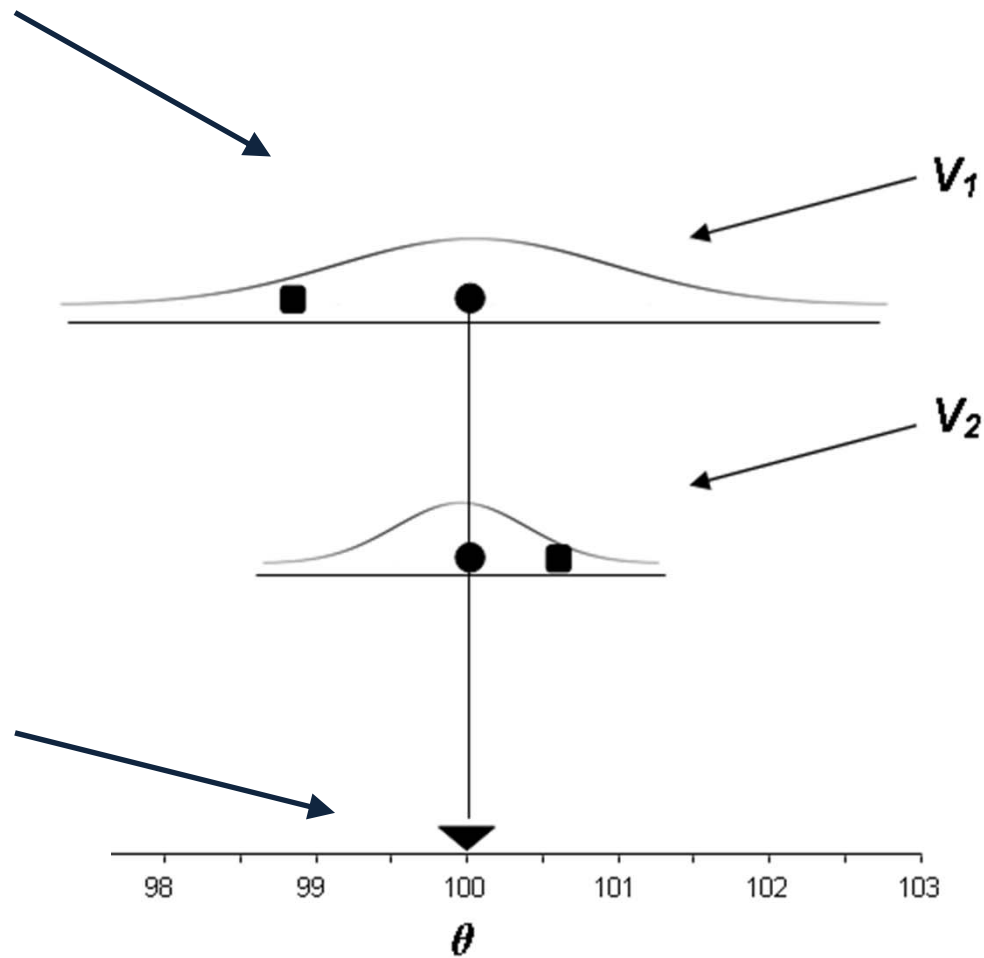
when studies share a common effect size



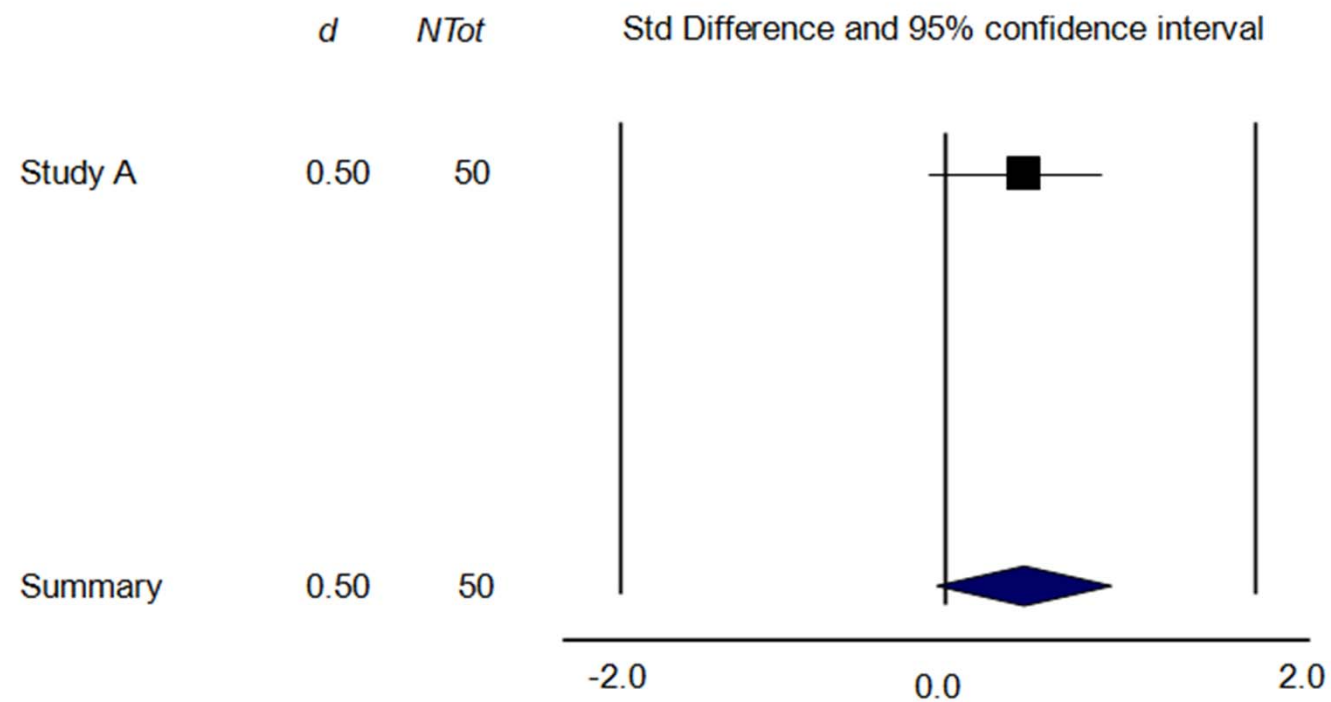
Weights

when studies share a common effect size

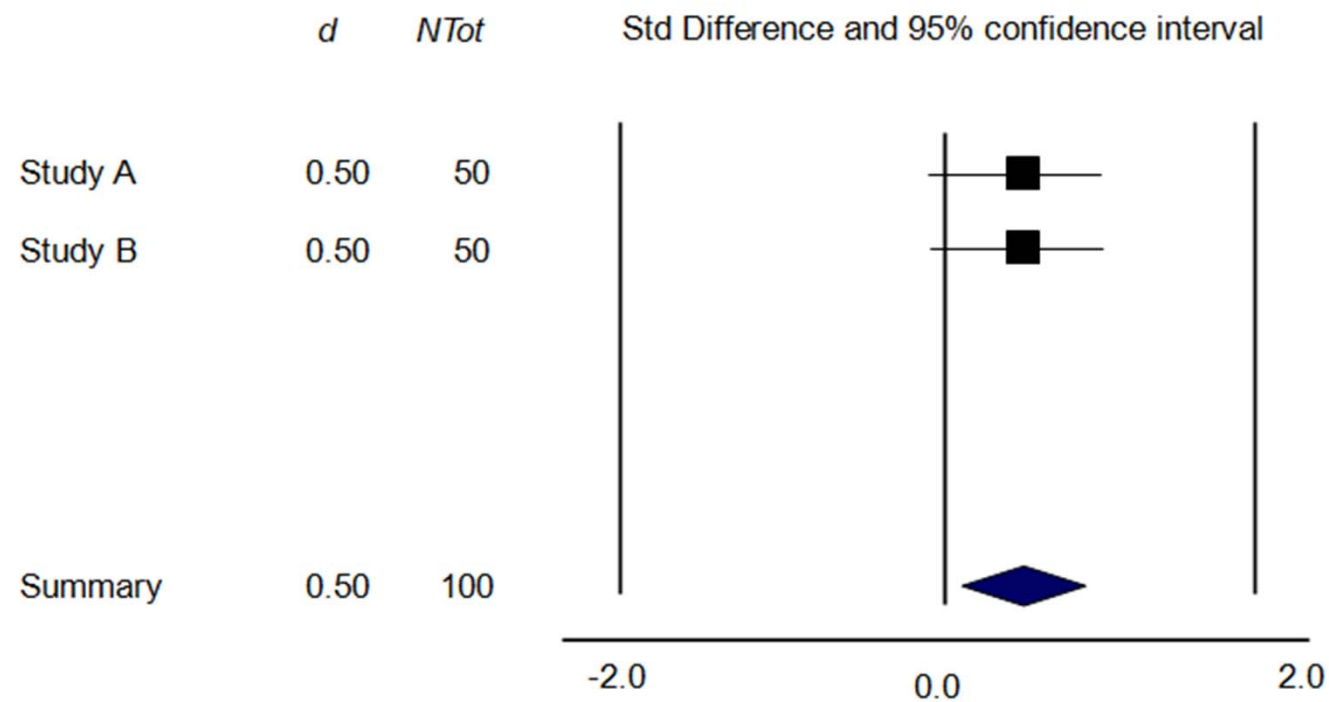
$$W = 1 / (V_1)$$



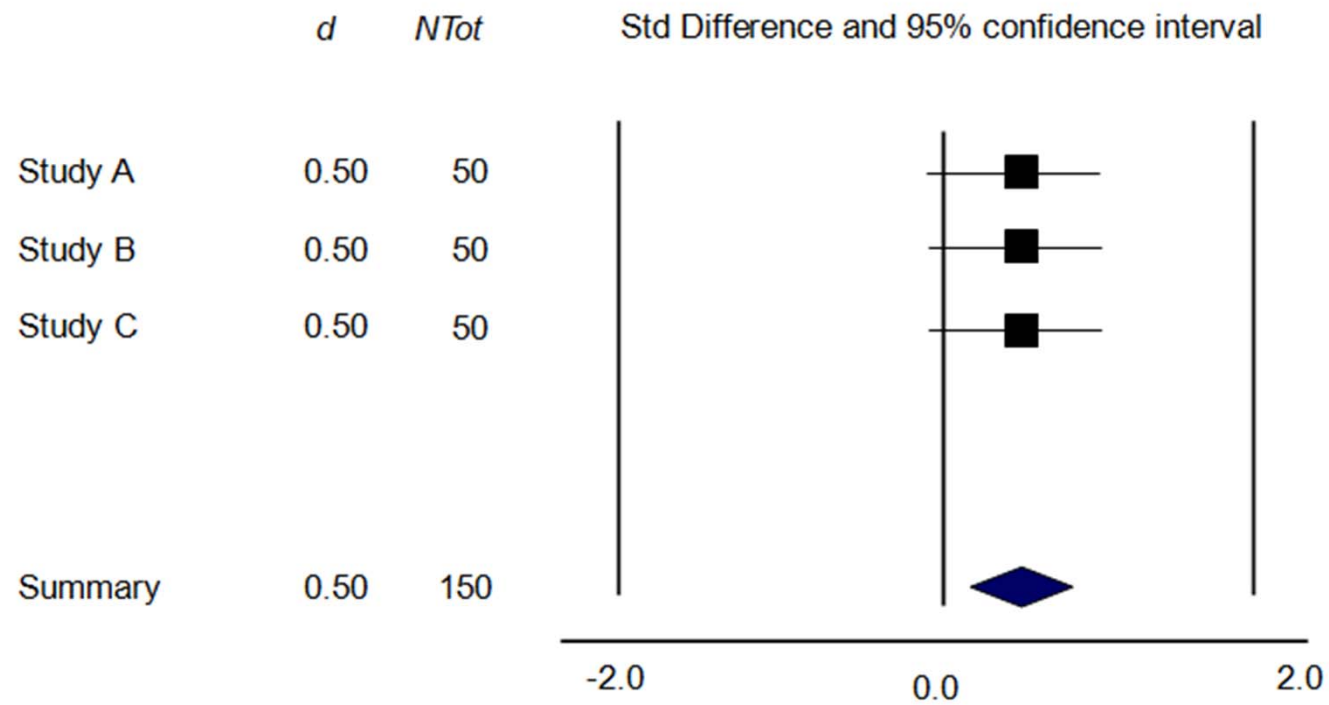
Meta-analysis with consistent effects $k = 1$



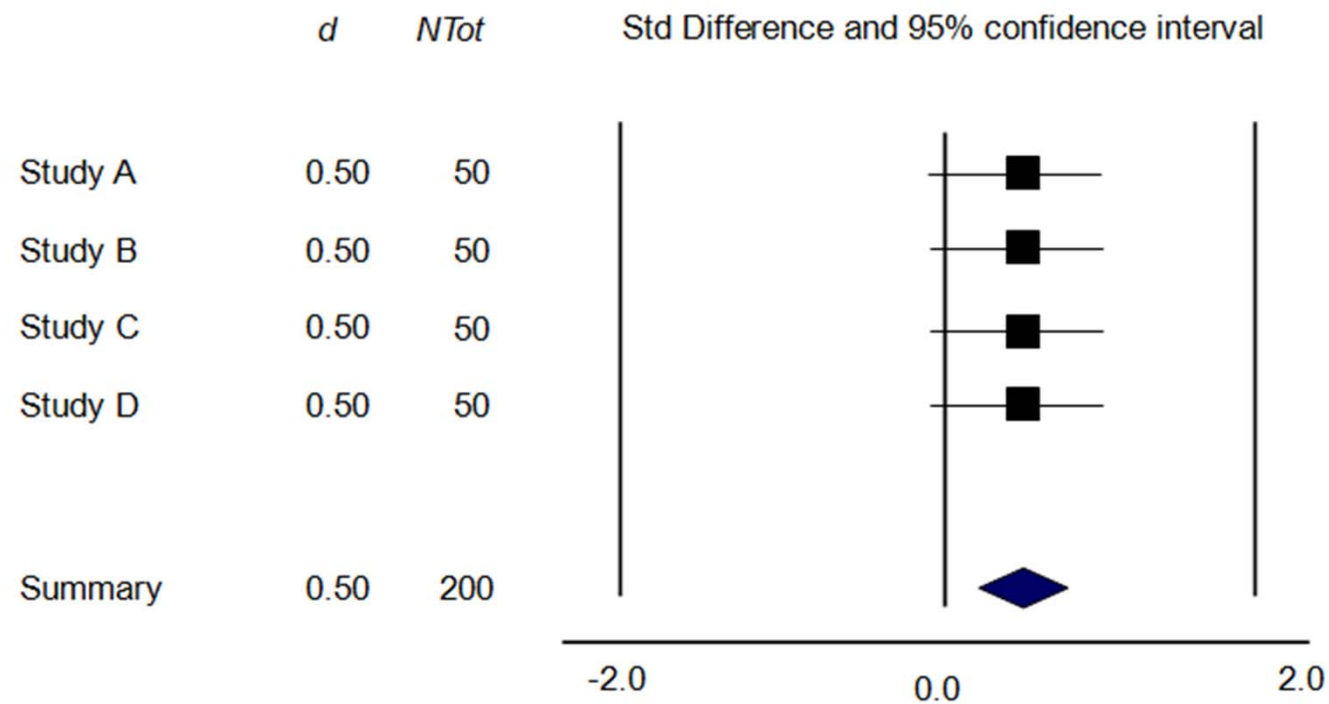
Meta-analysis with consistent effects $k = 2$



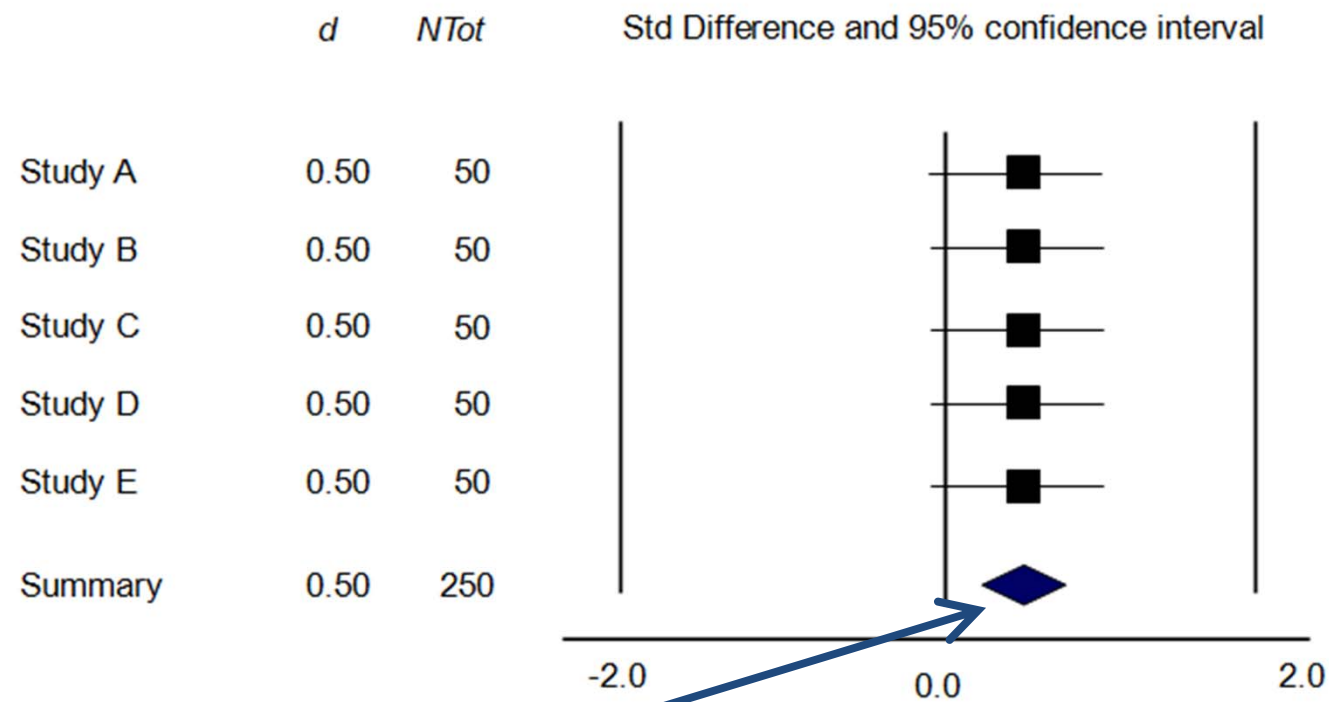
Meta-analysis with consistent effects $k = 3$



Meta-analysis with consistent effects $k = 4$



Meta-analysis with consistent effects $k = 5$



Precision of the
mean effect

The impact of weights

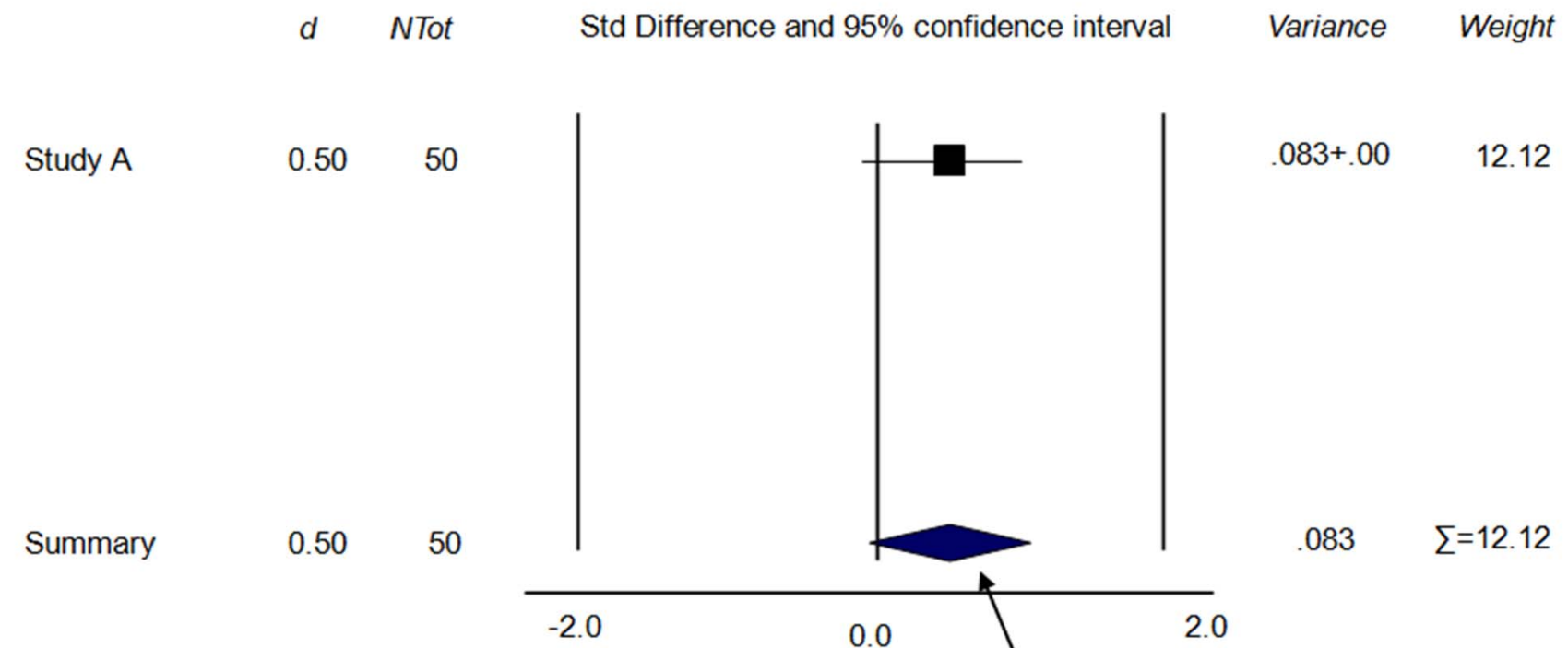
Weight = total amount of information

For single study $\text{Weight} = 1/V$

For combined effect $V = 1/(\text{Sum of Weights})$

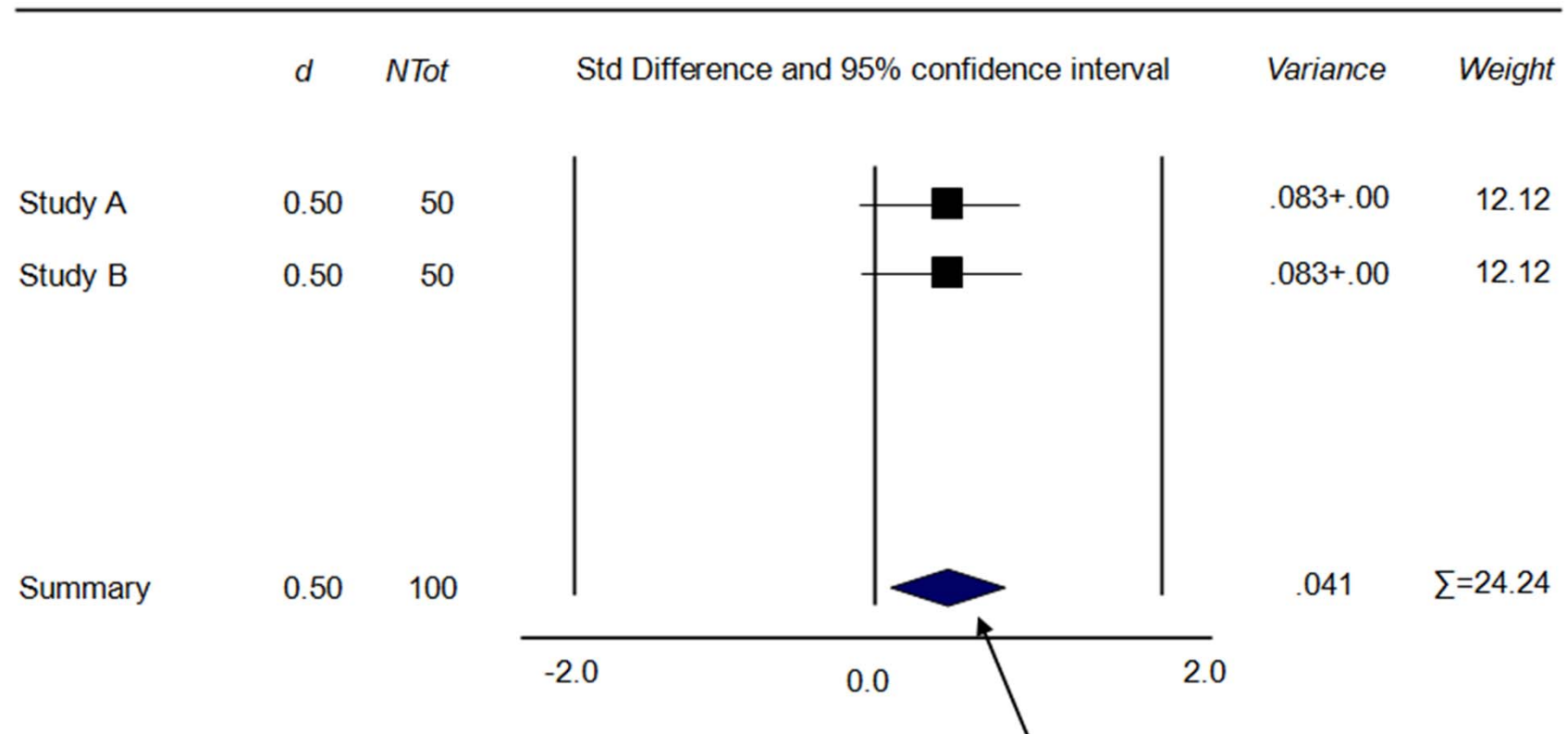
Impact on precision

Meta-analysis with consistent effects $k = 1$



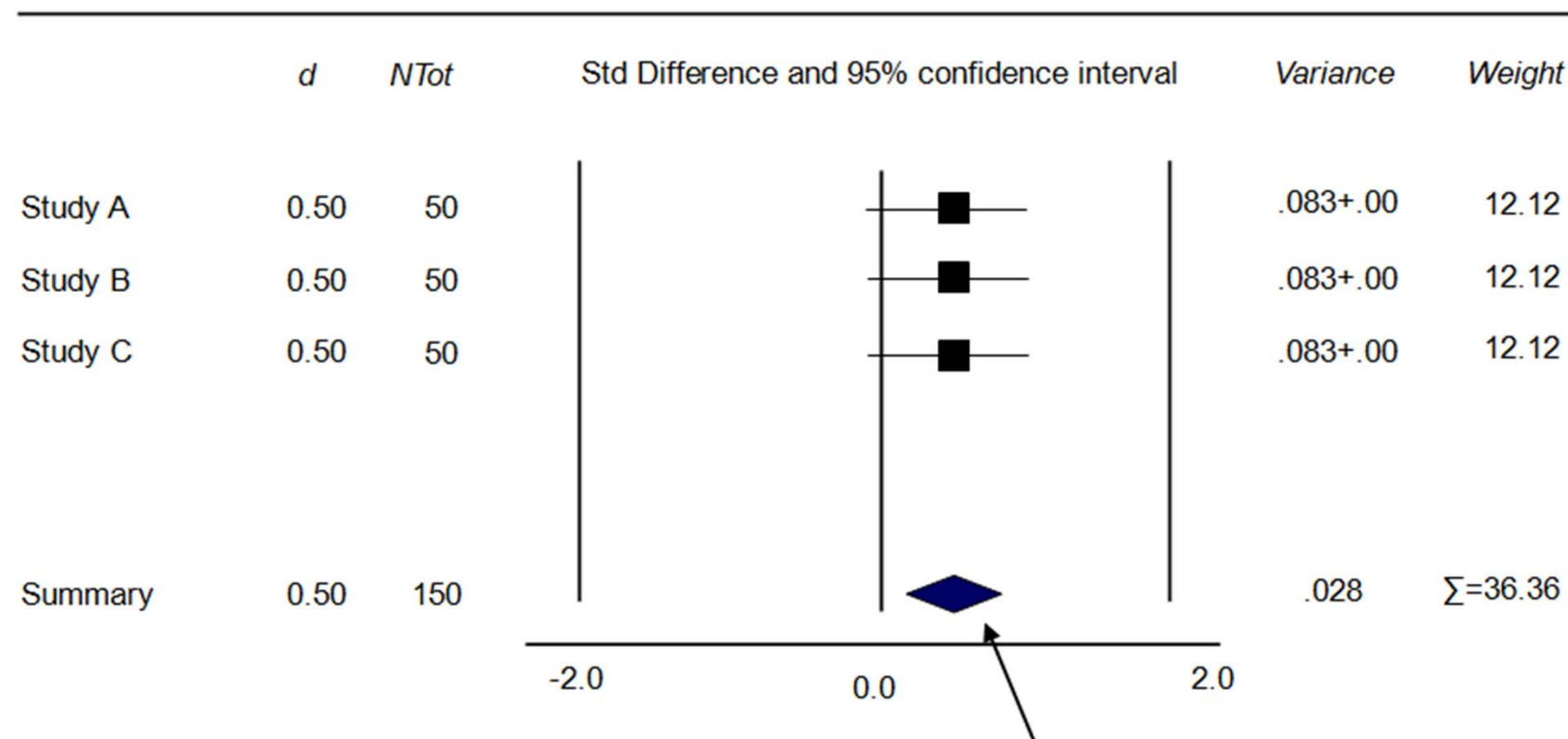
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{6.06}{12.12} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{12.12} = 0.083 \quad SE = \sqrt{0.083} = 0.287$$

Meta-analysis with consistent effects $k = 2$



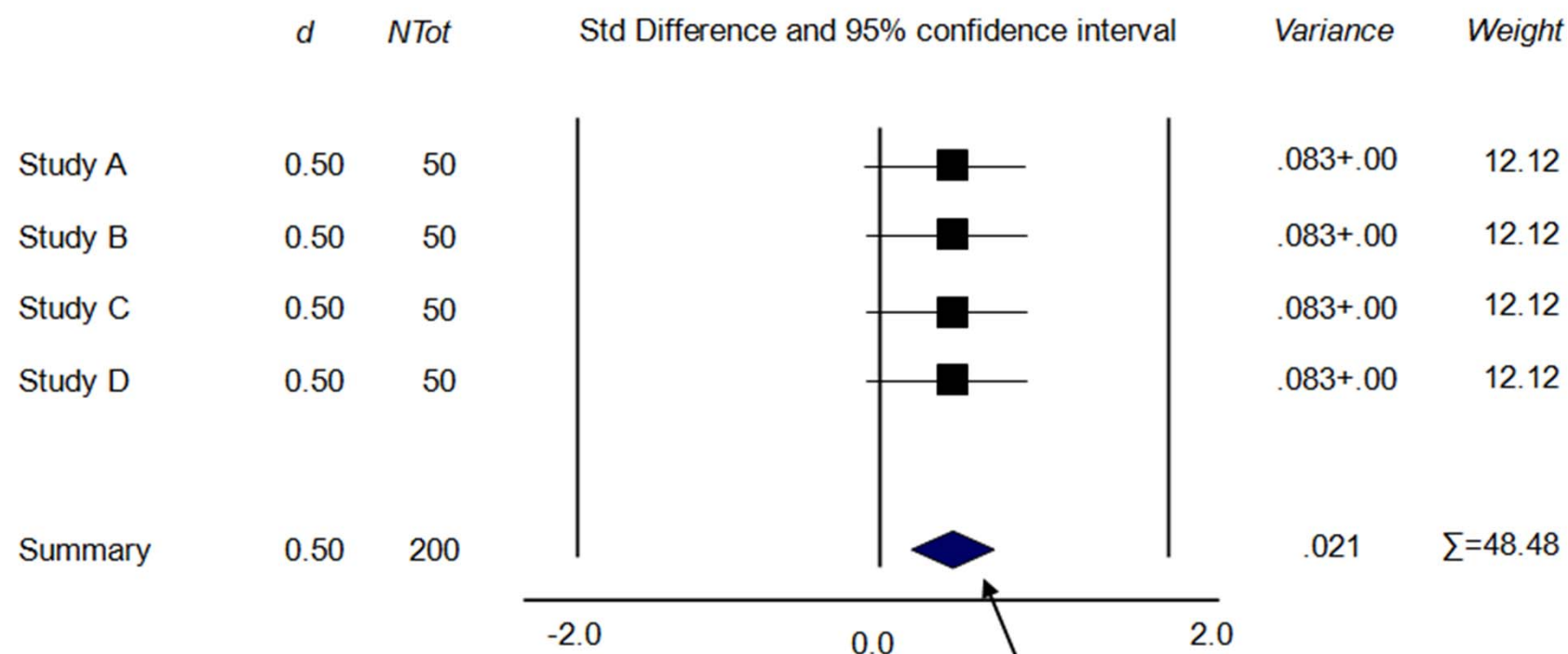
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{12.12}{24.24} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{24.24} = 0.041 \quad SE = \sqrt{0.041} = 0.203$$

Meta-analysis with consistent effects $k = 3$



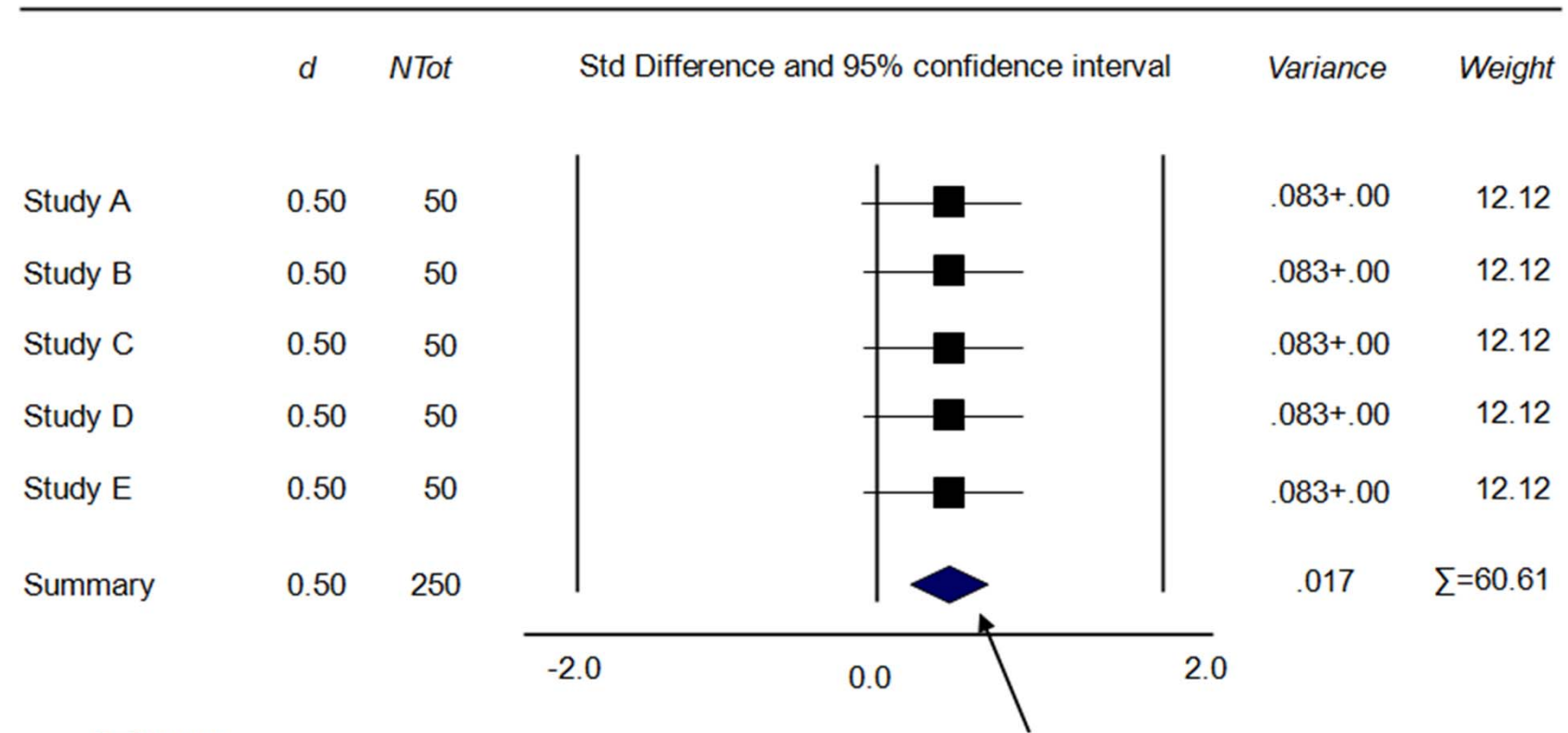
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{18.18}{36.36} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{36.36} = 0.028 \quad SE = \sqrt{0.028} = 0.166$$

Meta-analysis with consistent effects $k = 4$



$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{24.24}{48.48} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{48.48} = 0.021 \quad SE = \sqrt{0.021} = 0.144$$

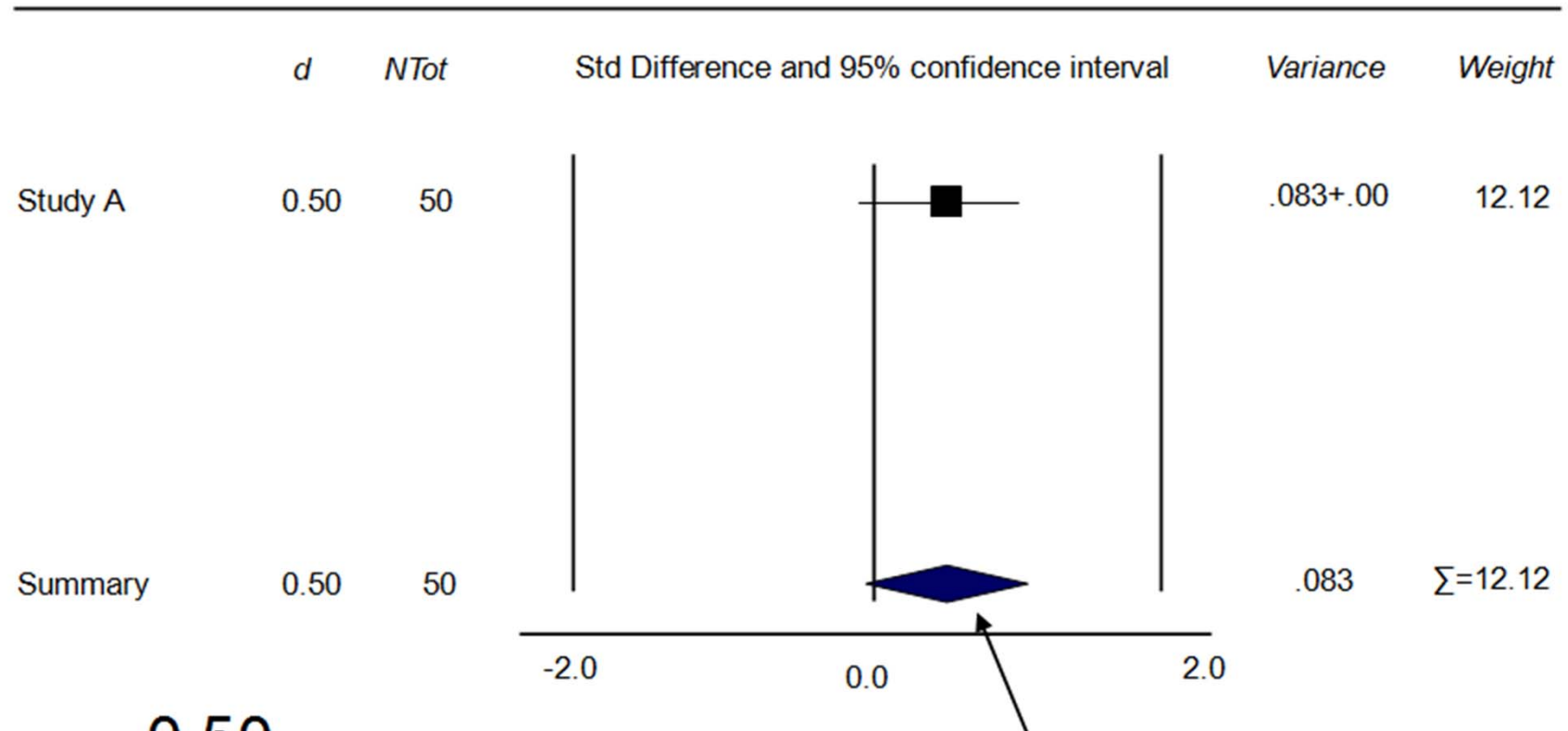
Meta-analysis with consistent effects $k = 5$



$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{30.30}{60.60} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{60.61} = 0.017 \quad SE = \sqrt{0.017} = 0.128$$

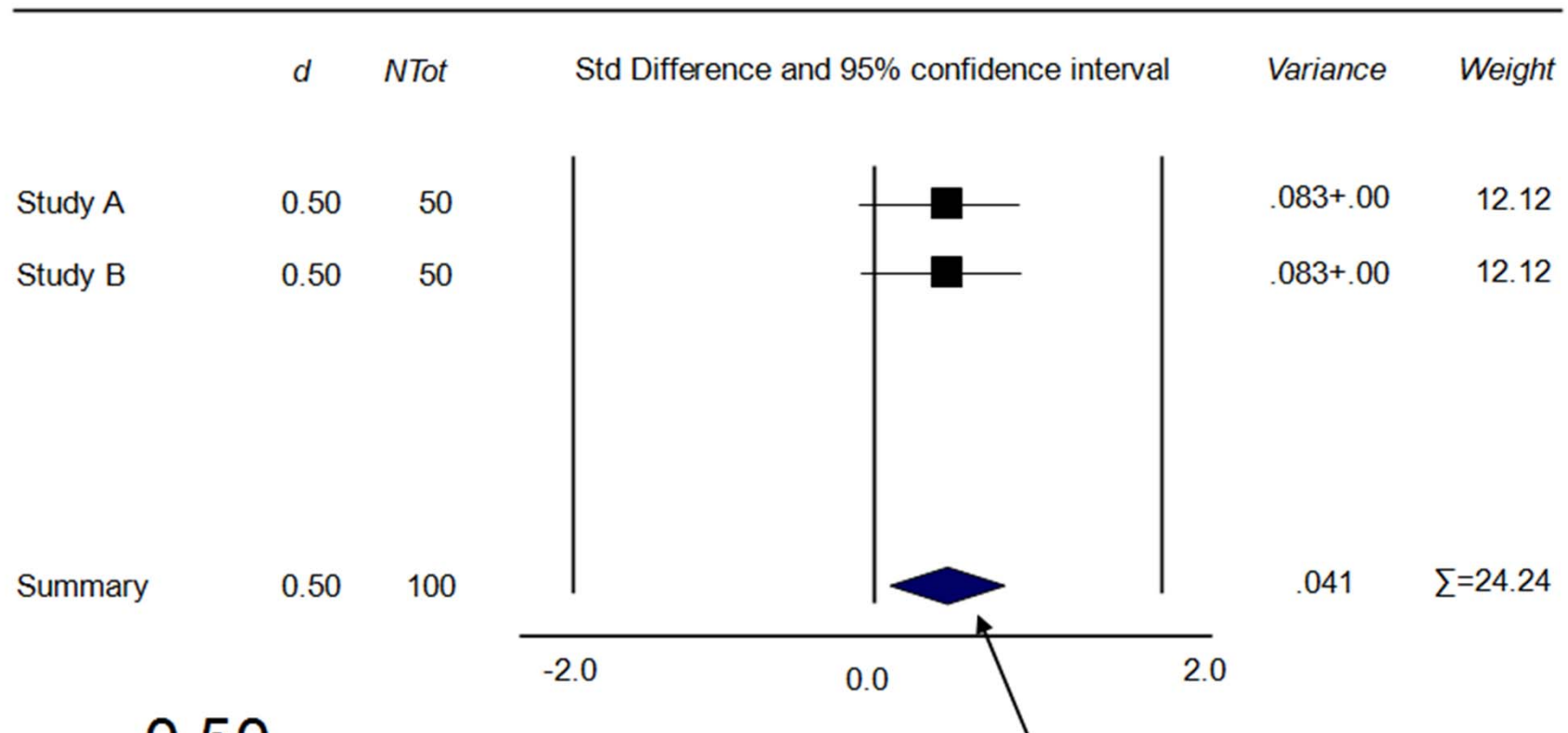
Impact on p -value

Meta-analysis with consistent effects $k = 1$



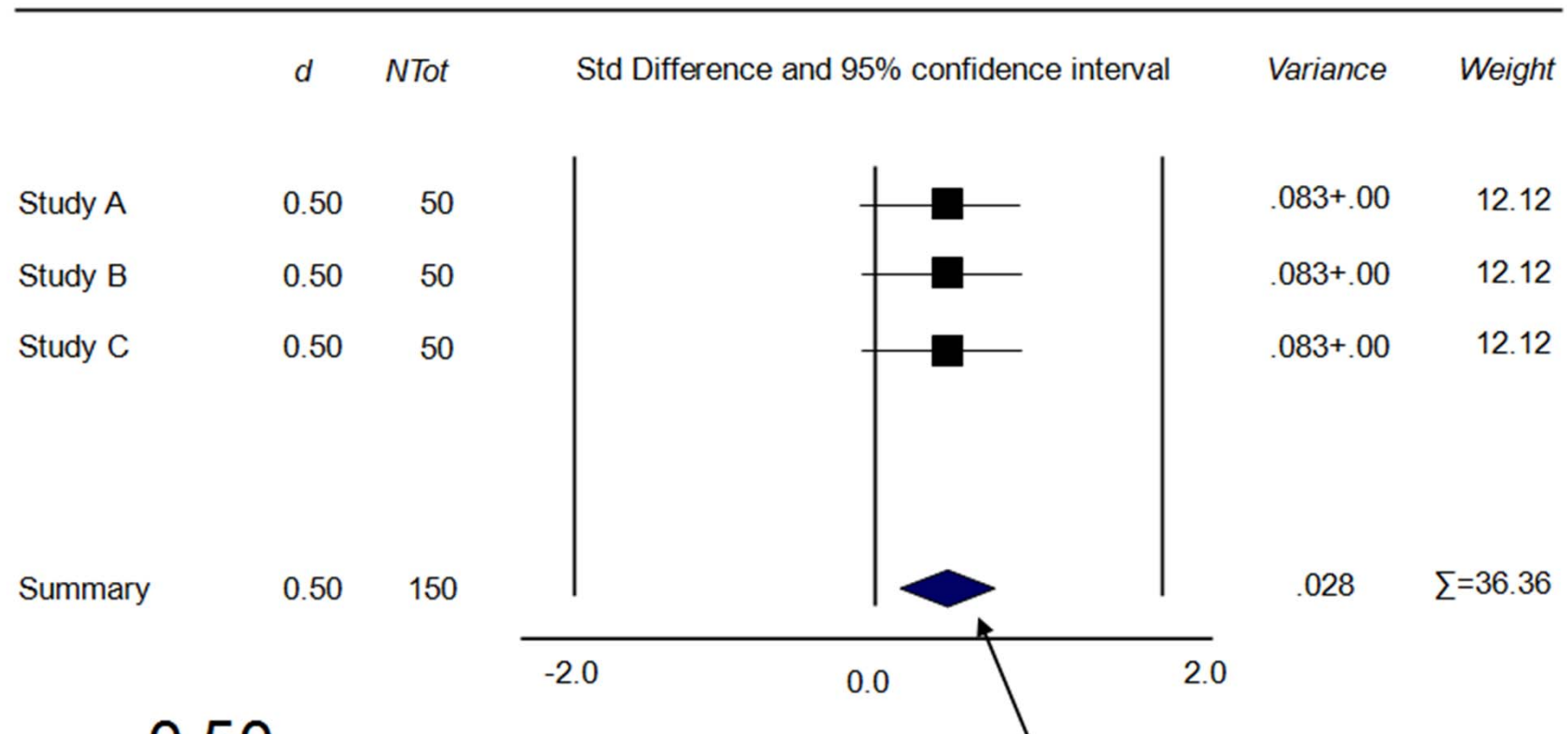
$$Z = \frac{0.50}{.287} = 1.74, p = .082$$

Meta-analysis with consistent effects $k = 2$



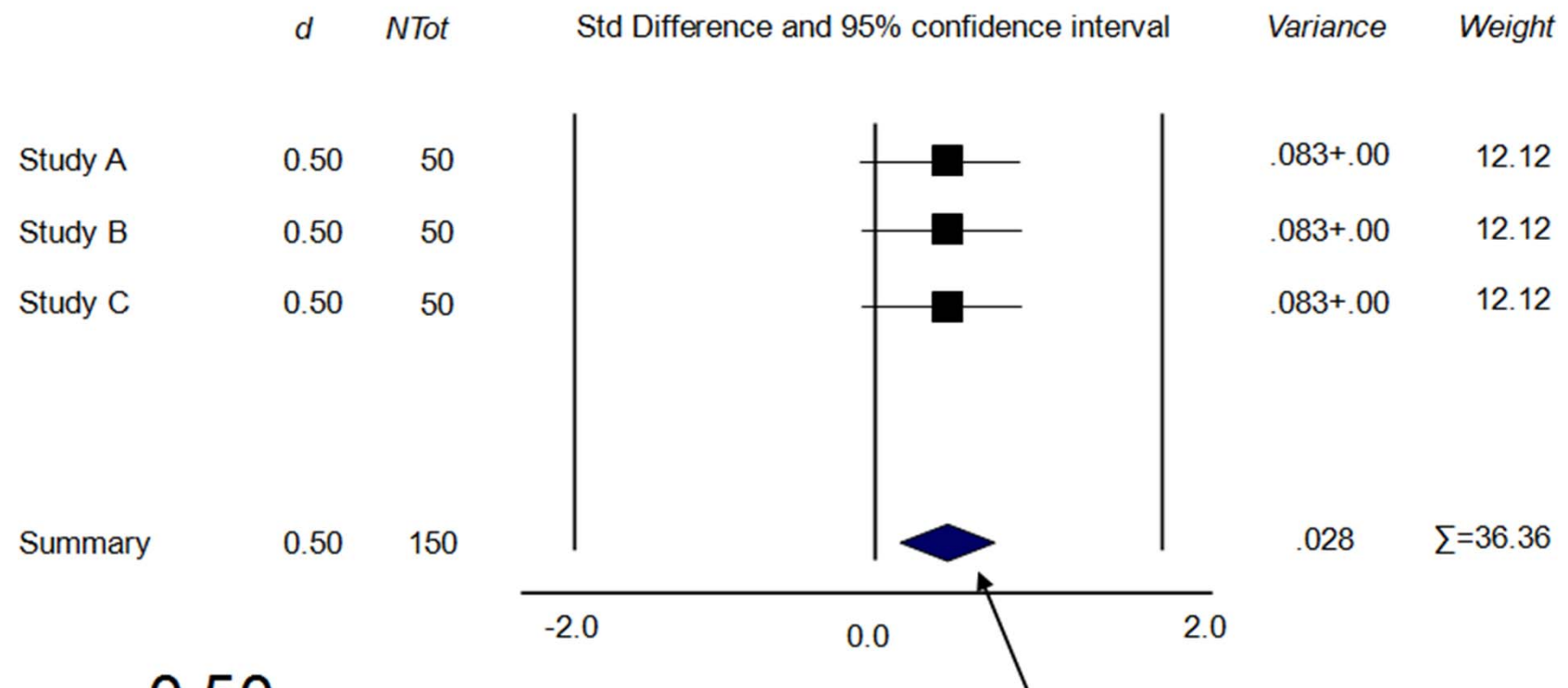
$$Z = \frac{0.50}{.203} = 2.46, p = .014$$

Meta-analysis with consistent effects $k = 3$



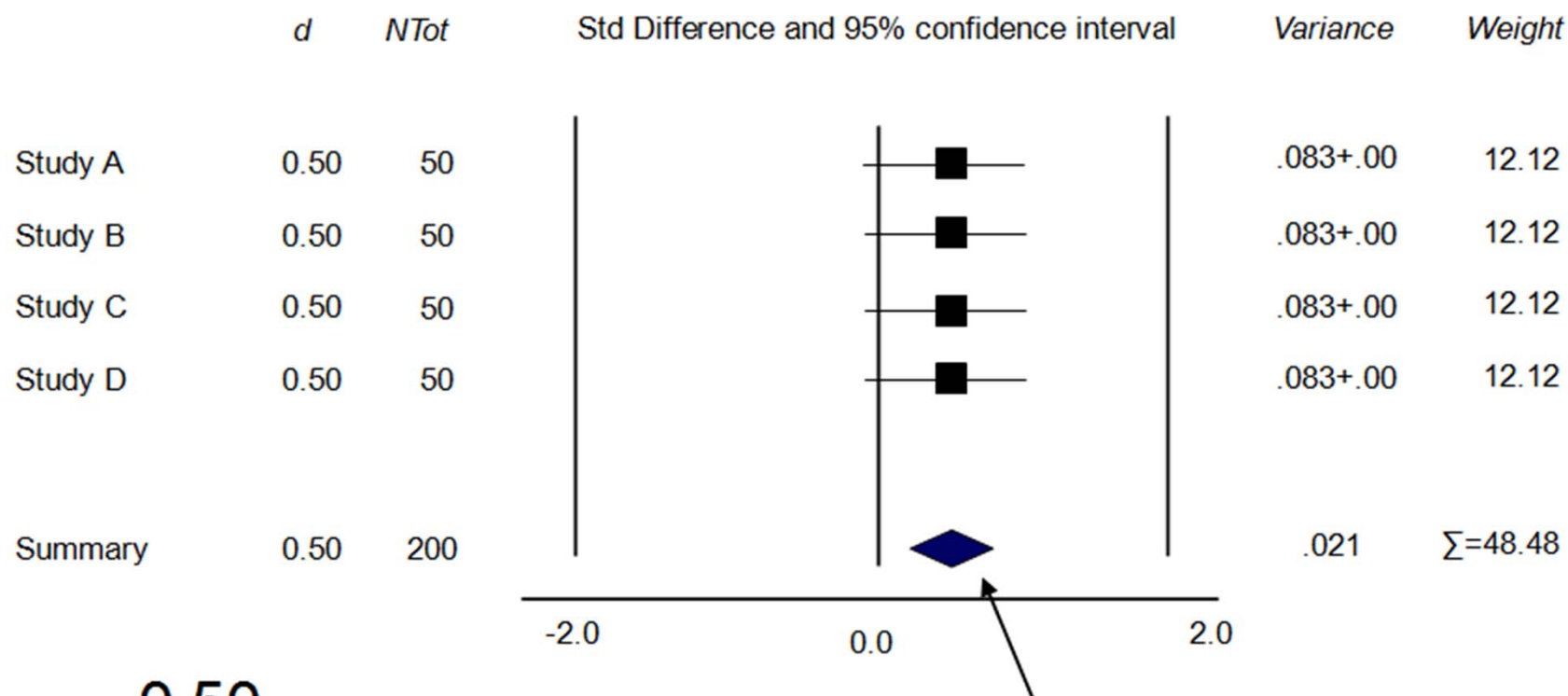
$$Z = \frac{0.50}{.166} = 3.02, p = .003$$

Meta-analysis with consistent effects $k = 3$



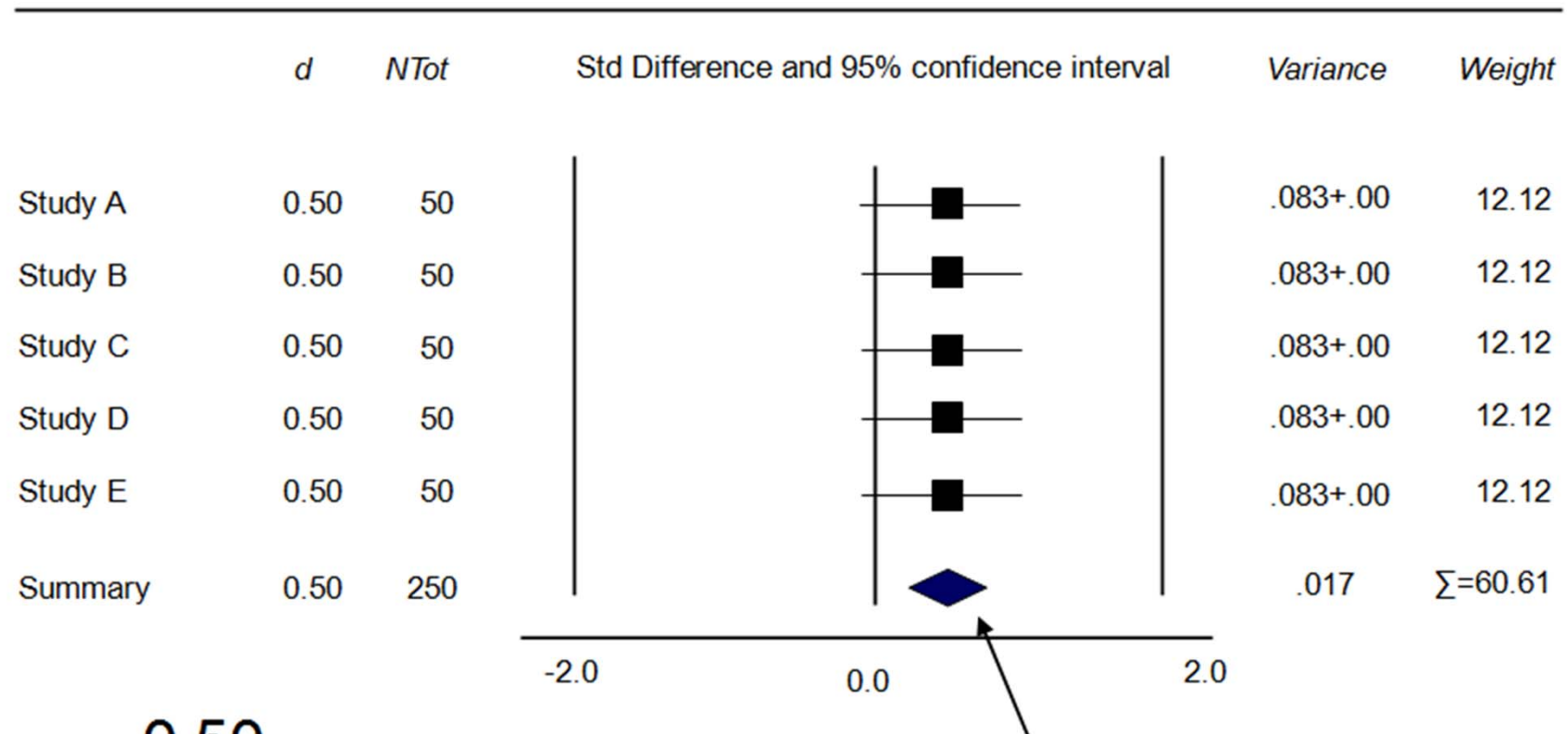
$$Z = \frac{0.50}{.166} = 3.02, p = .003$$

Meta-analysis with consistent effects $k = 4$



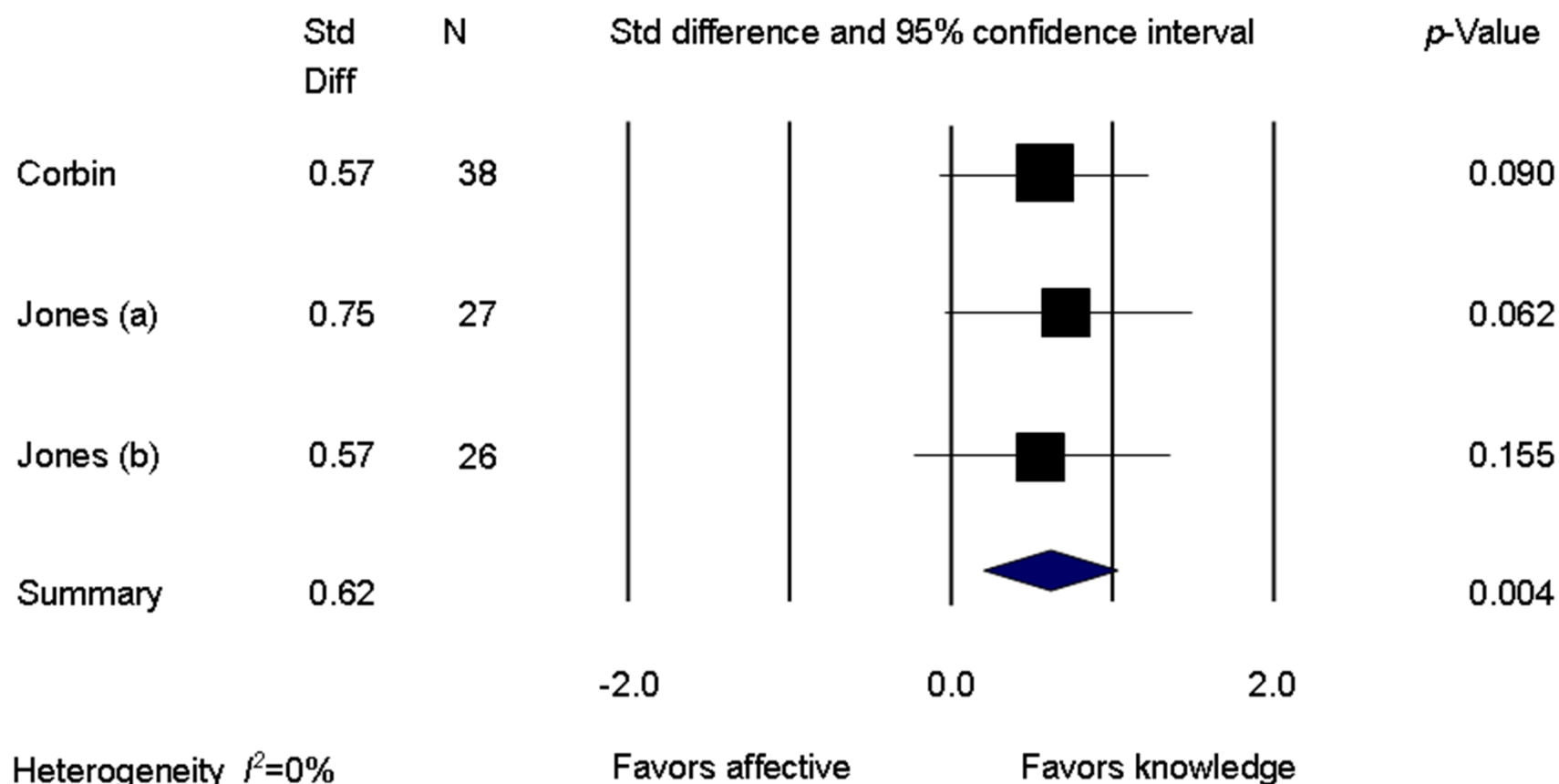
$$Z = \frac{0.50}{.144} = 3.48, p = .0005$$

Meta-analysis with consistent effects $k = 5$



$$Z = \frac{0.50}{.128} = 3.89, p = .0001$$

Affective vs. knowledge-based training Drug knowledge

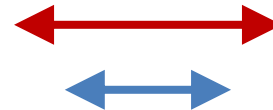
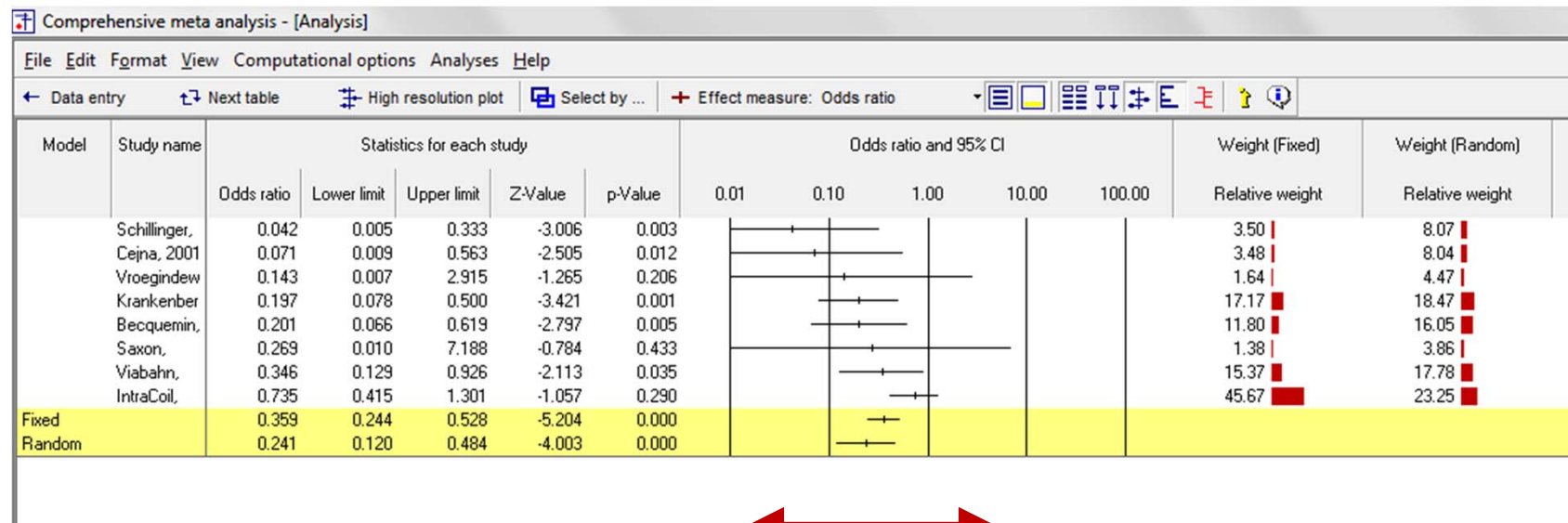


Heterogeneity $I^2=0\%$

$Q=0.1$, $df=2$, $p=0.93$

Heterogeneity in effect sizes

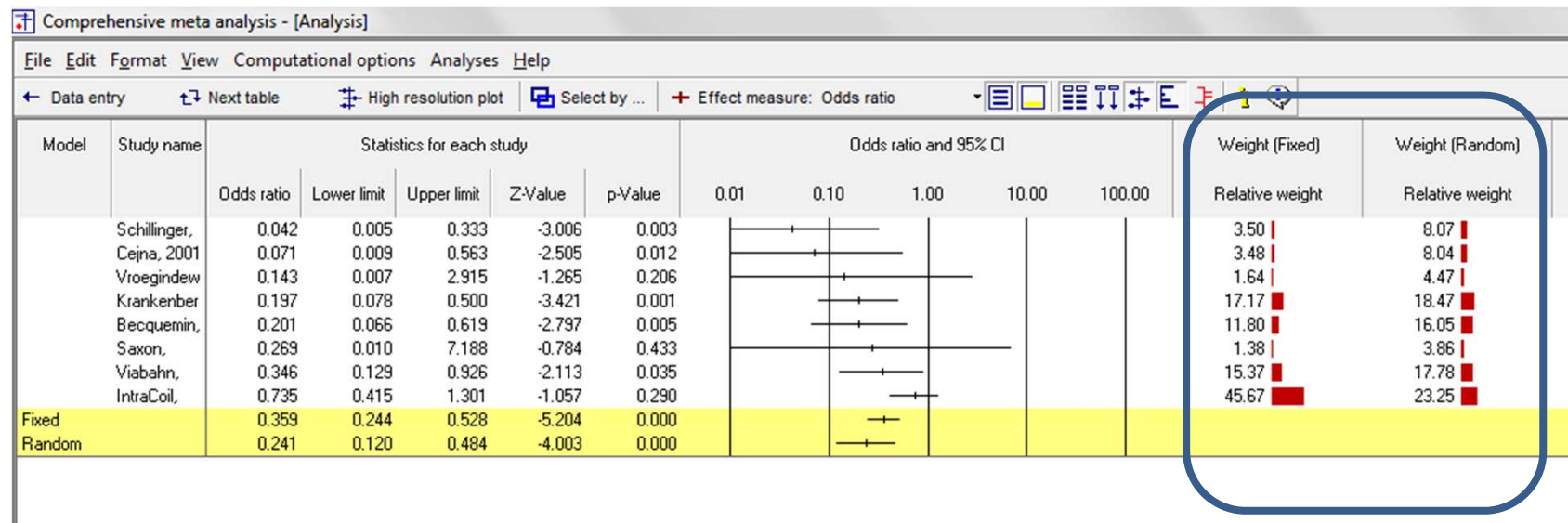
What we mean by Heterogeneity



It's the variance in *true effects* (not observed effects) that we care about

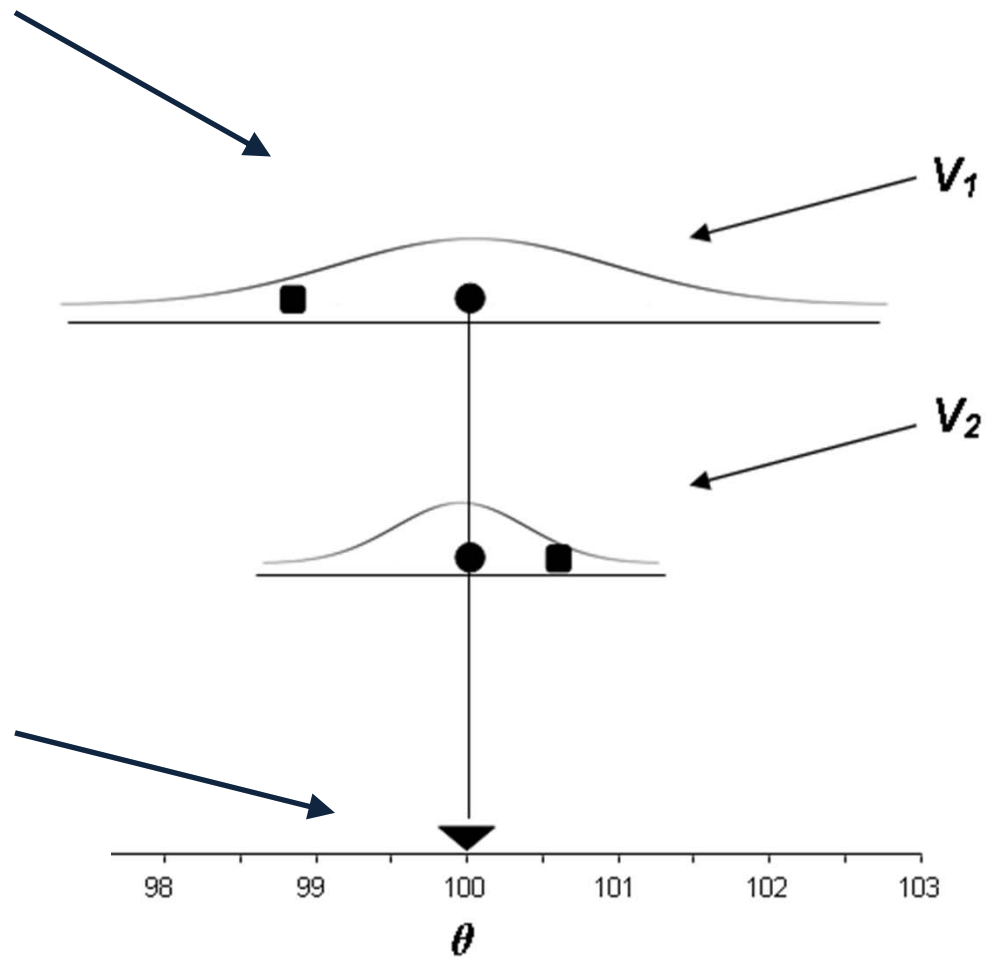
Why estimate heterogeneity?

It affects the weights



Weights when $T^2 = 0$

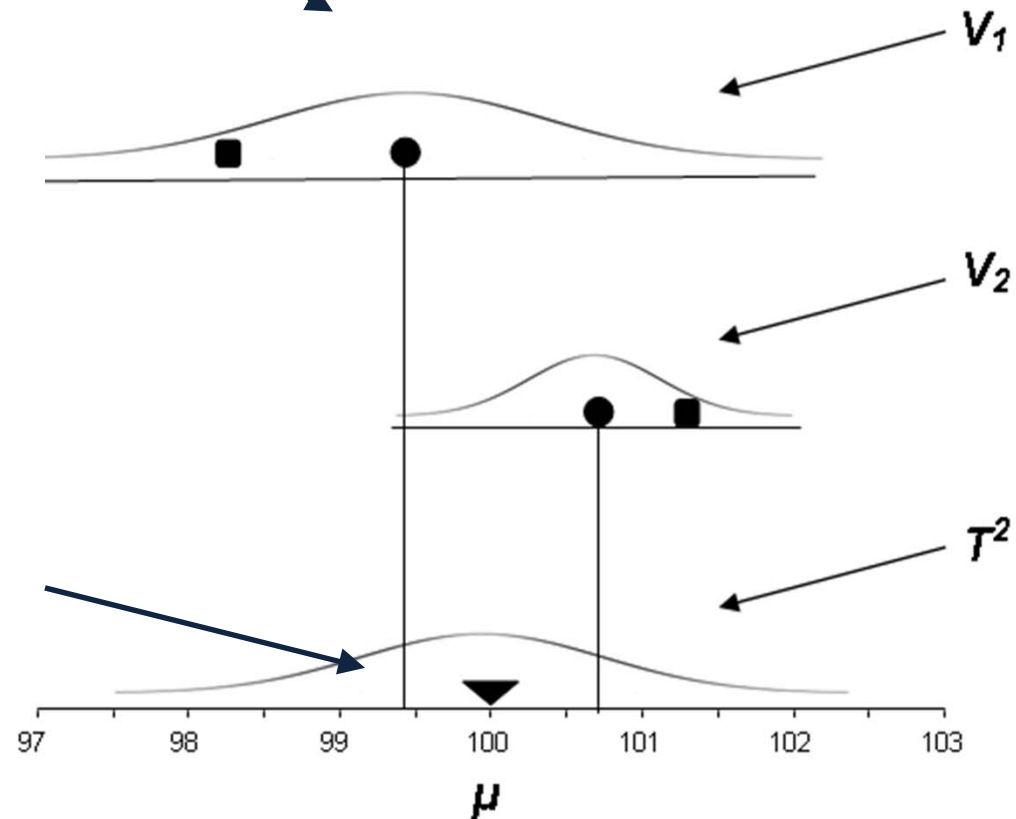
$$W = 1 / (V_1)$$



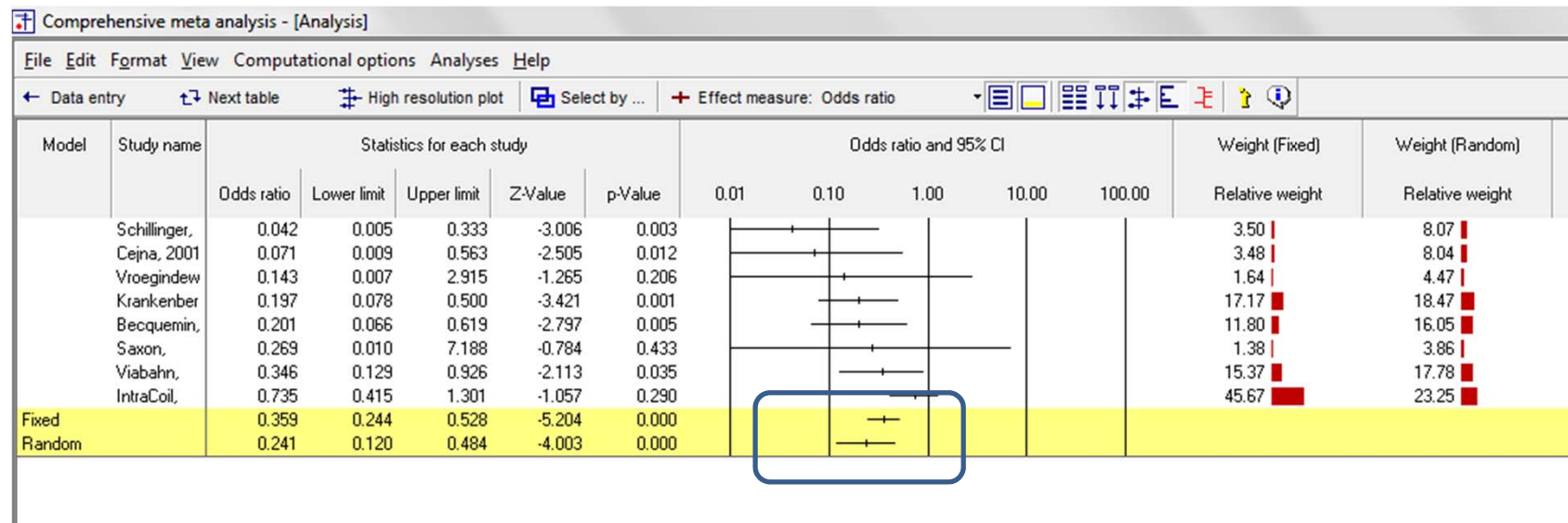
Weights when $T^2 > 0$

$$W = 1 / (V_1 + T^2)$$

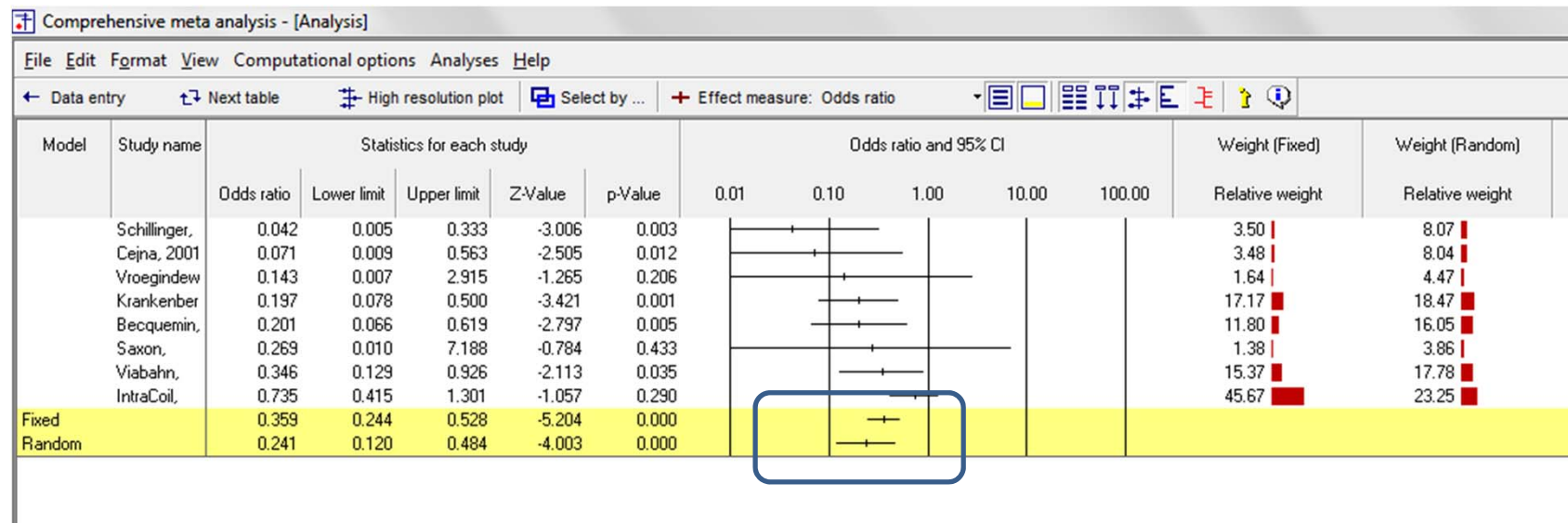
Study 2



It affects the mean

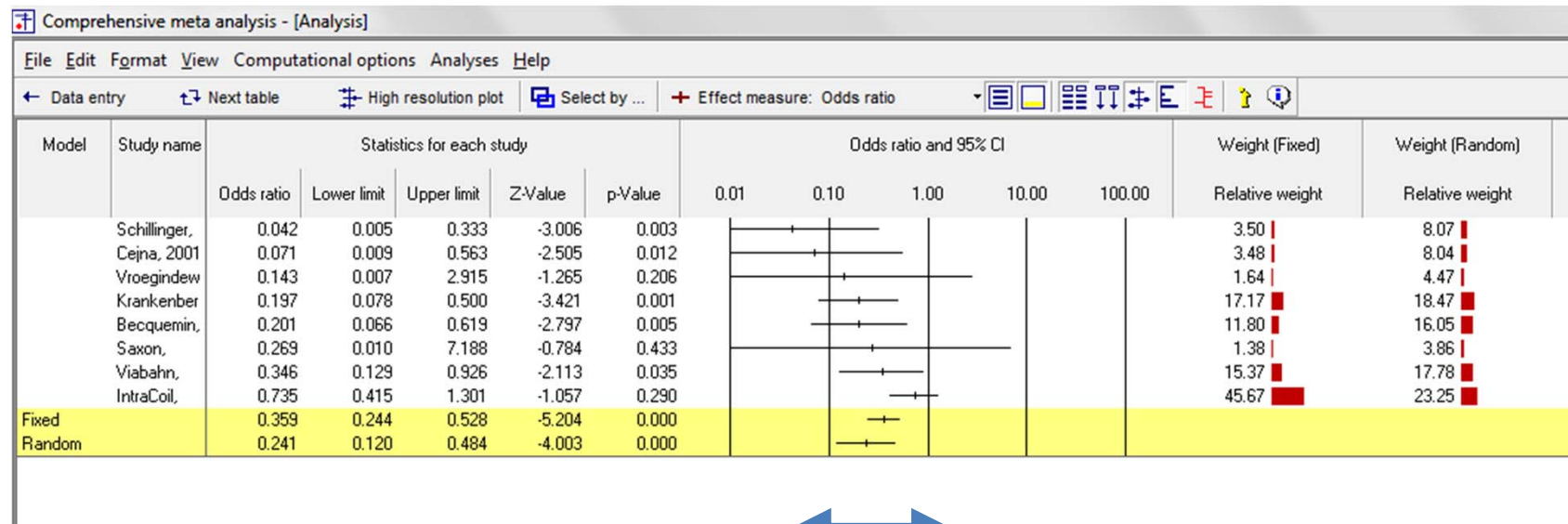


It affects the standard error ...



... the confidence interval, and the p-value

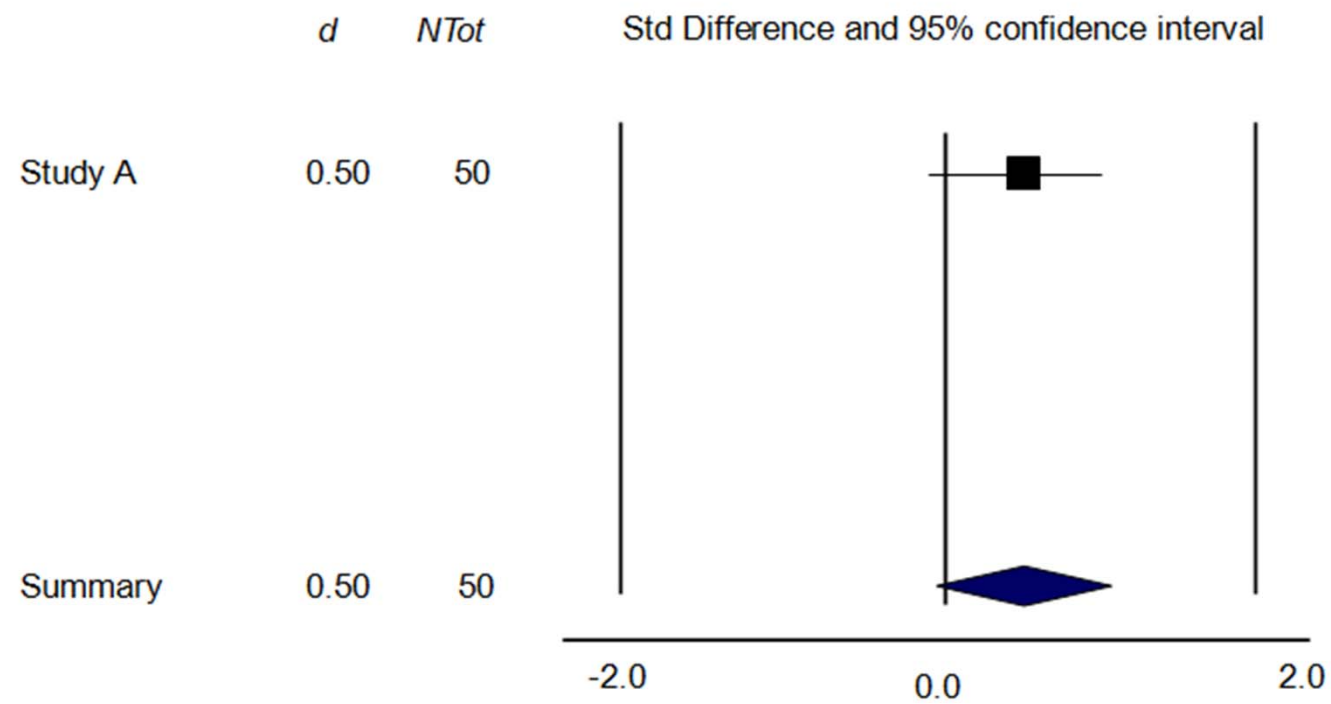
It affects the utility of the treatment



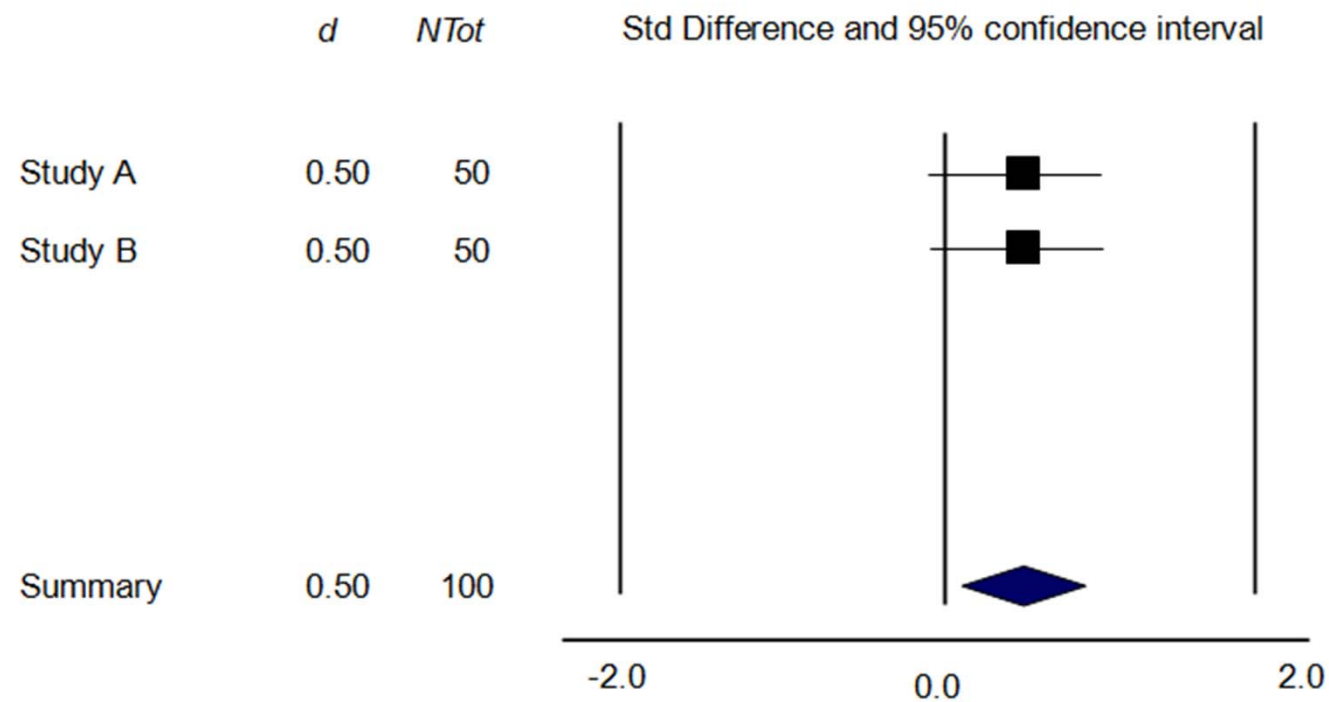
Is the treatment effective for everyone, or effective for some and harmful for others?

Impact on Standard Error

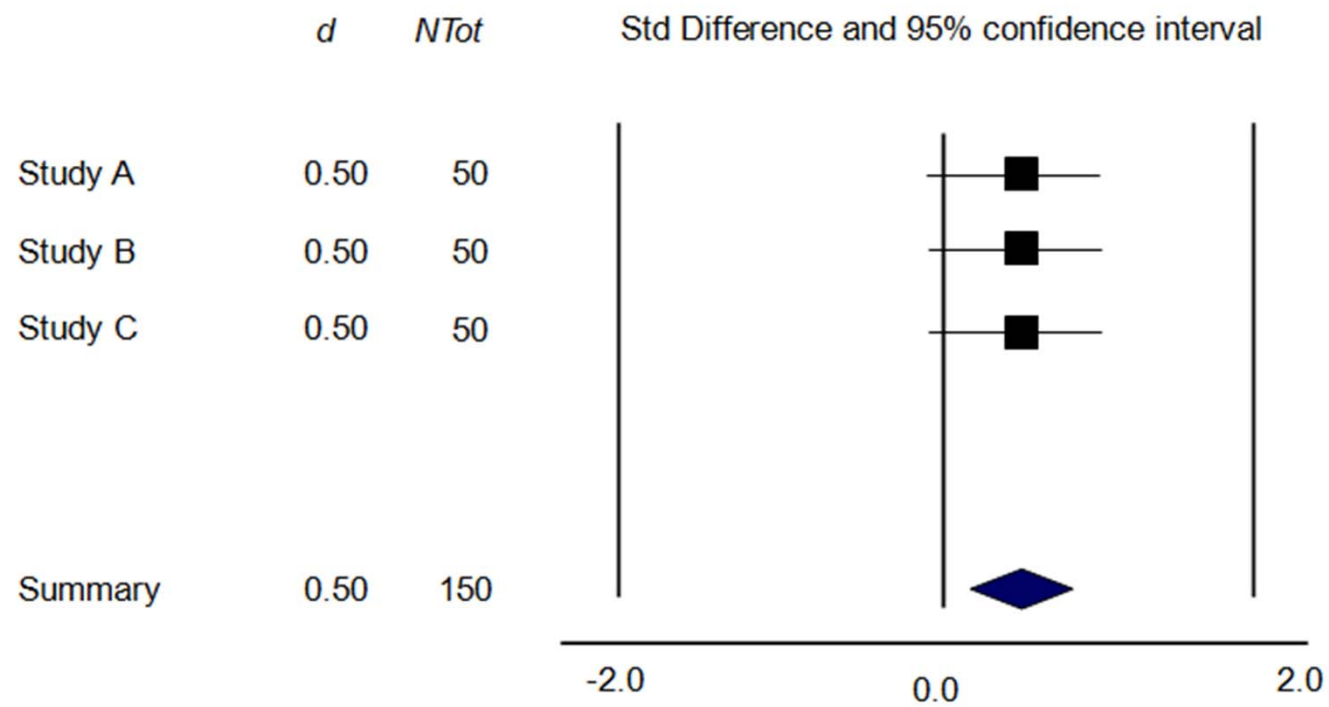
Meta-analysis with consistent effects $k = 1$



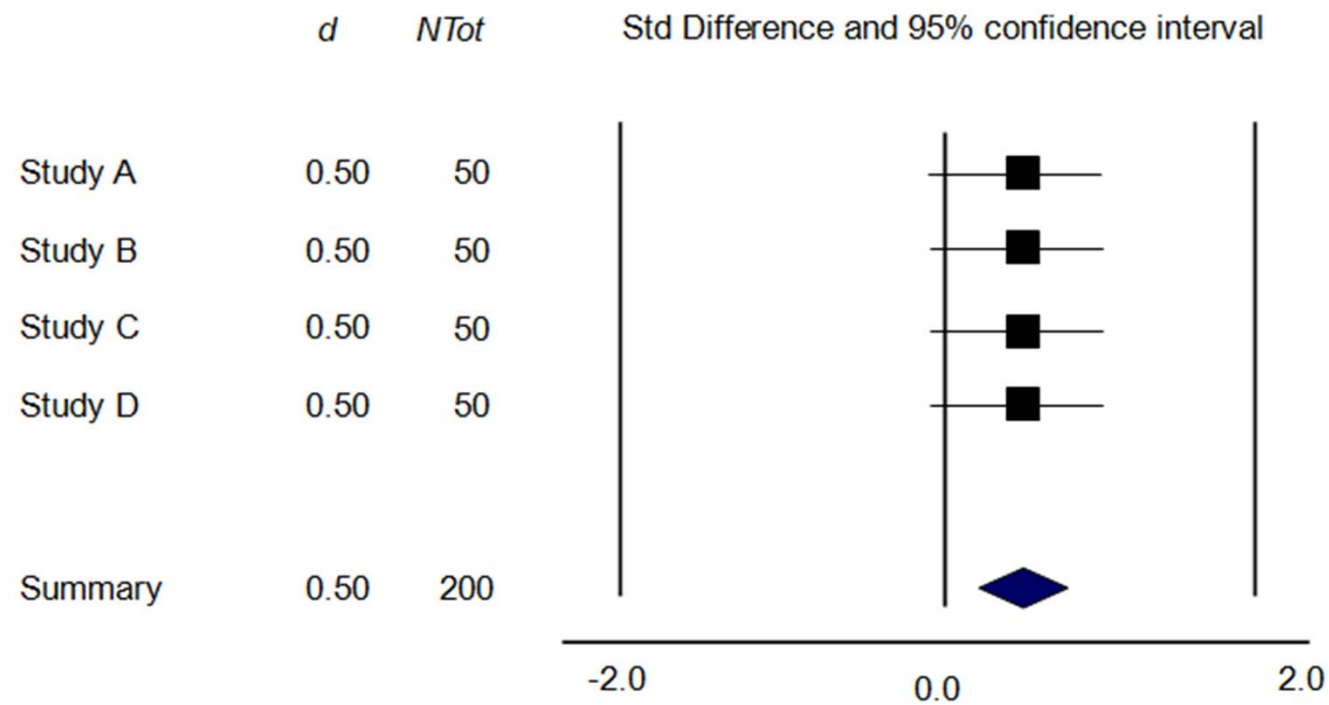
Meta-analysis with consistent effects $k = 2$



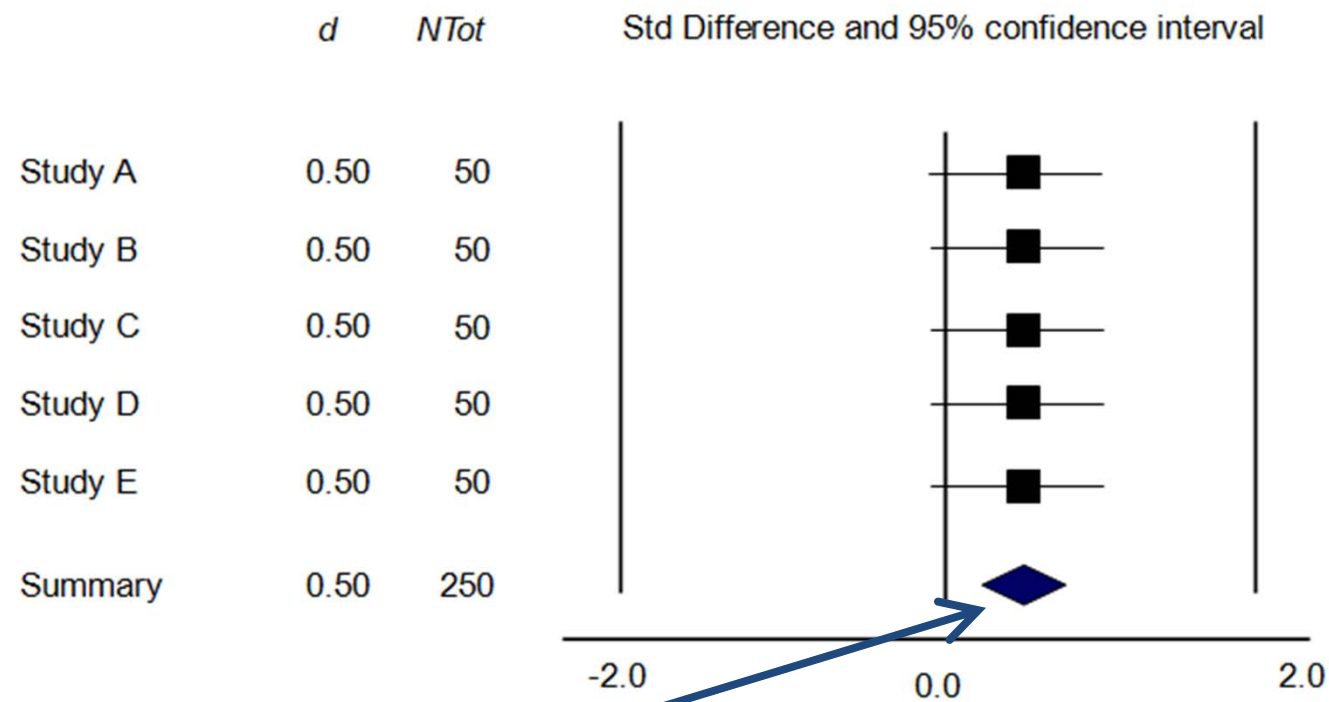
Meta-analysis with consistent effects $k = 3$



Meta-analysis with consistent effects $k = 4$

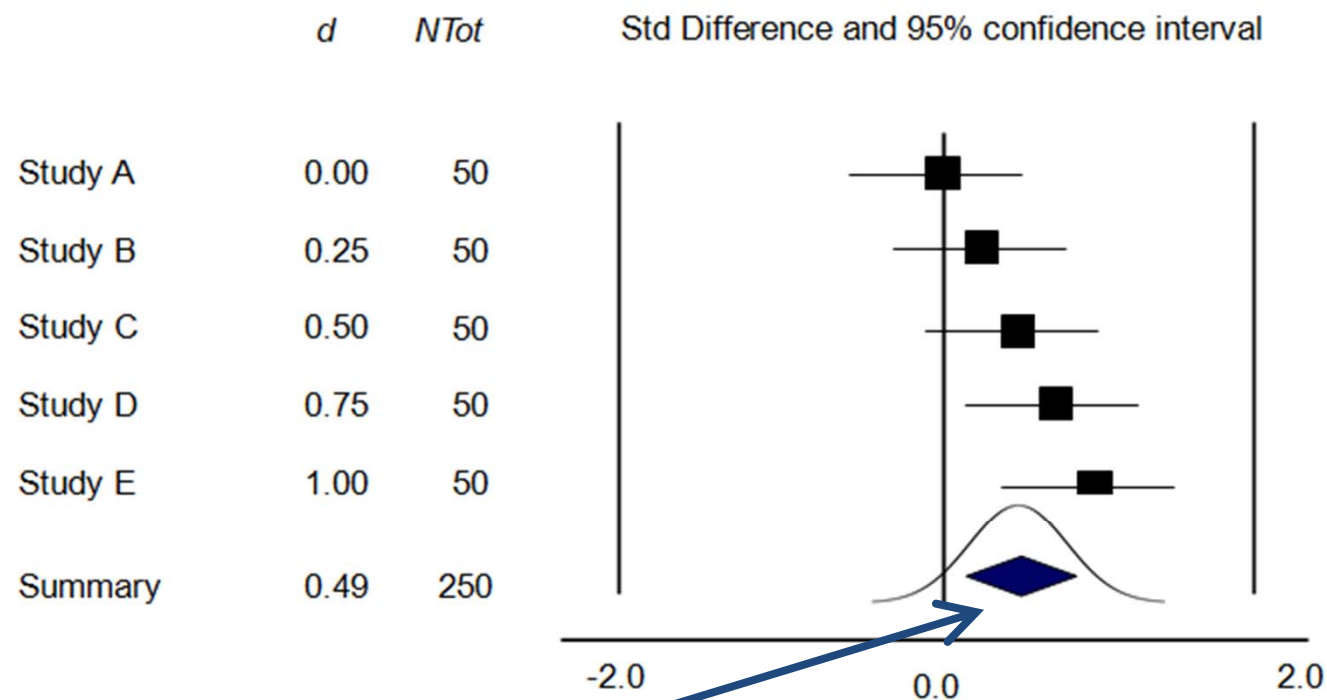


Meta-analysis with consistent effects $k = 5$



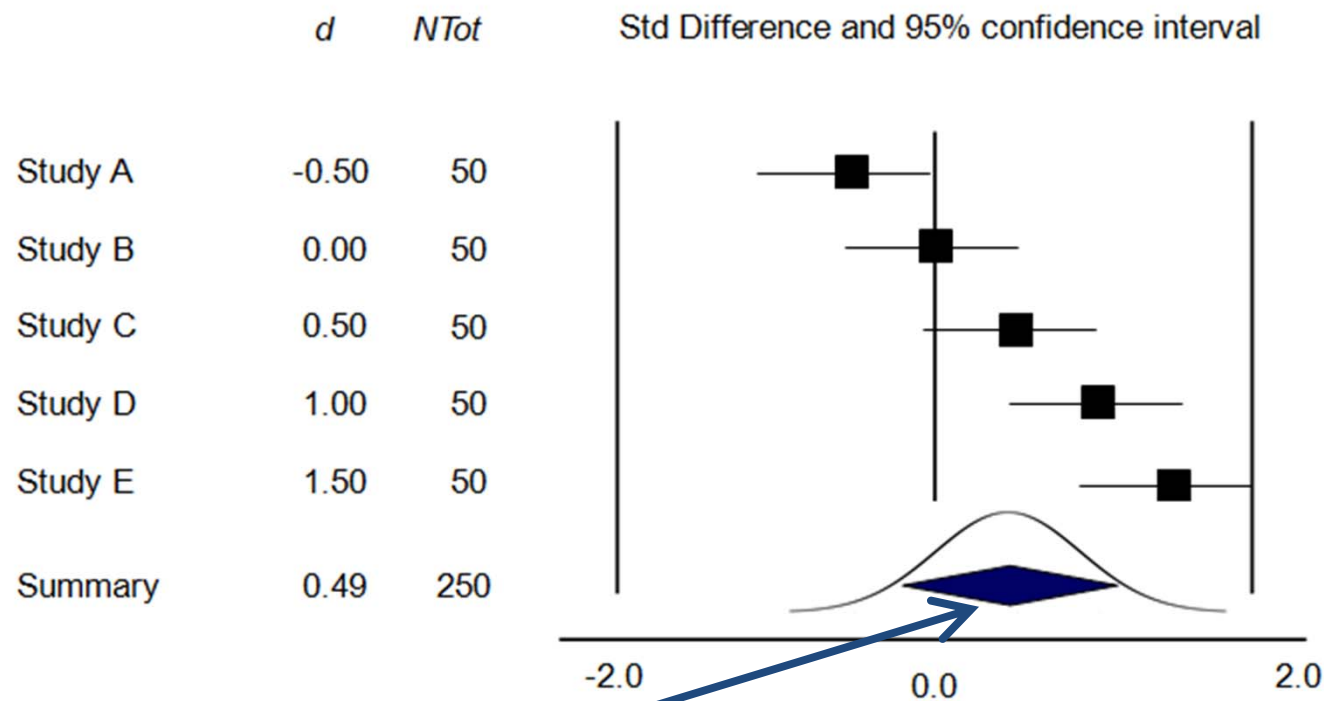
Precision of the
mean effect

Meta-analysis with heterogeneous effects $k = 5$



Precision of the
mean effect

Meta-analysis with heterogeneous effects $k = 5$

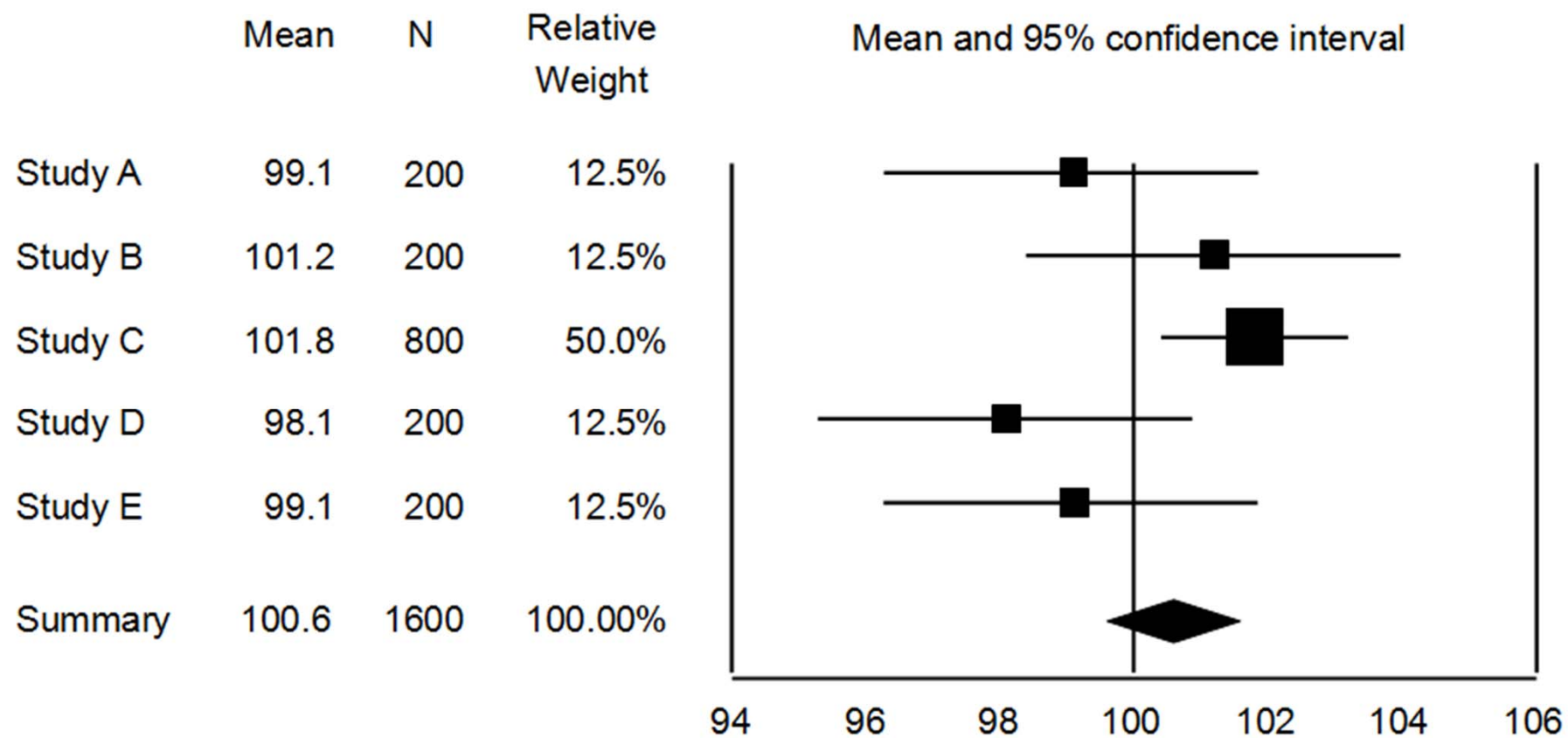


Precision of the
mean effect

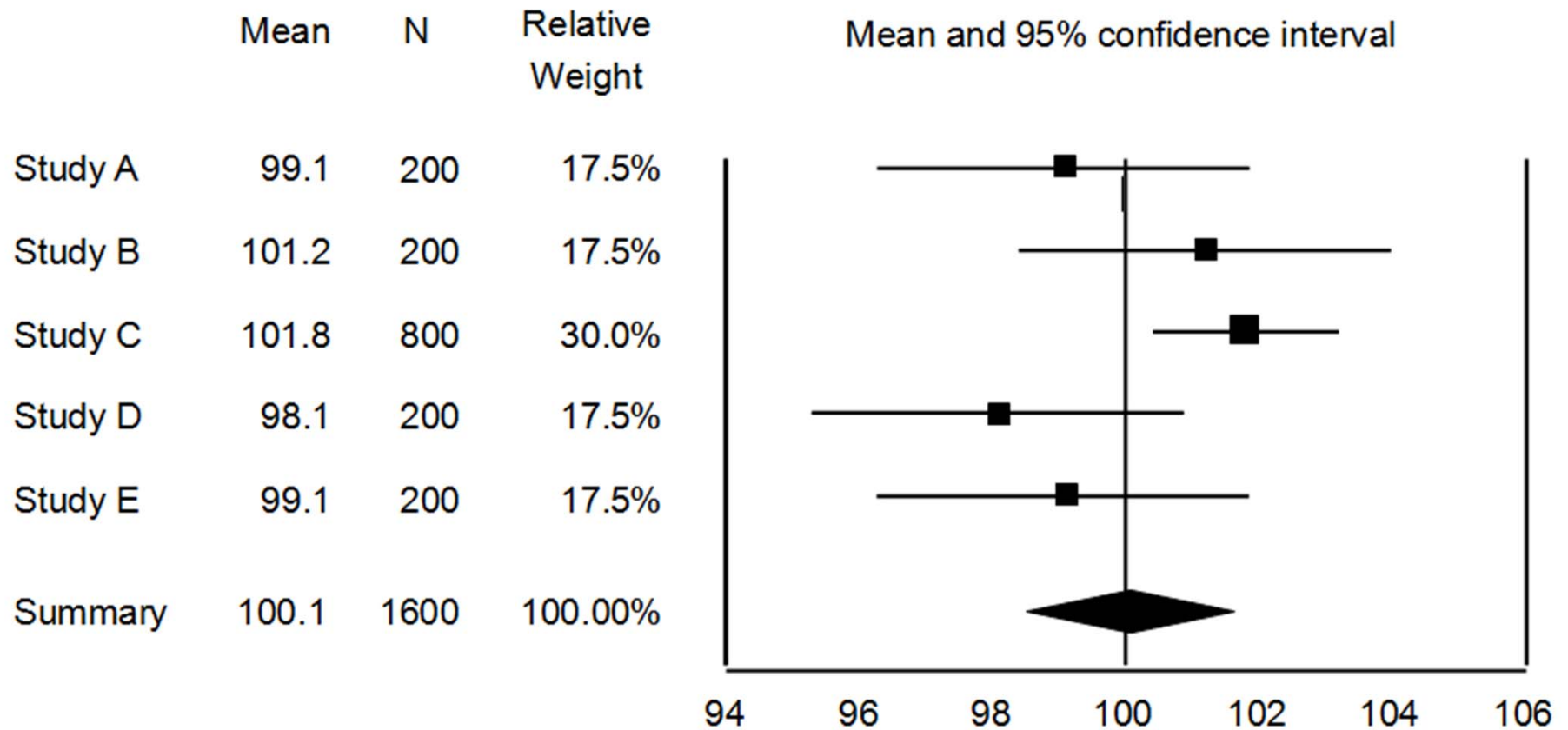
Impact on mean

- As heterogeneity increases, mean effect shifts away from larger studies and towards smaller studies

Aptitude score at one college

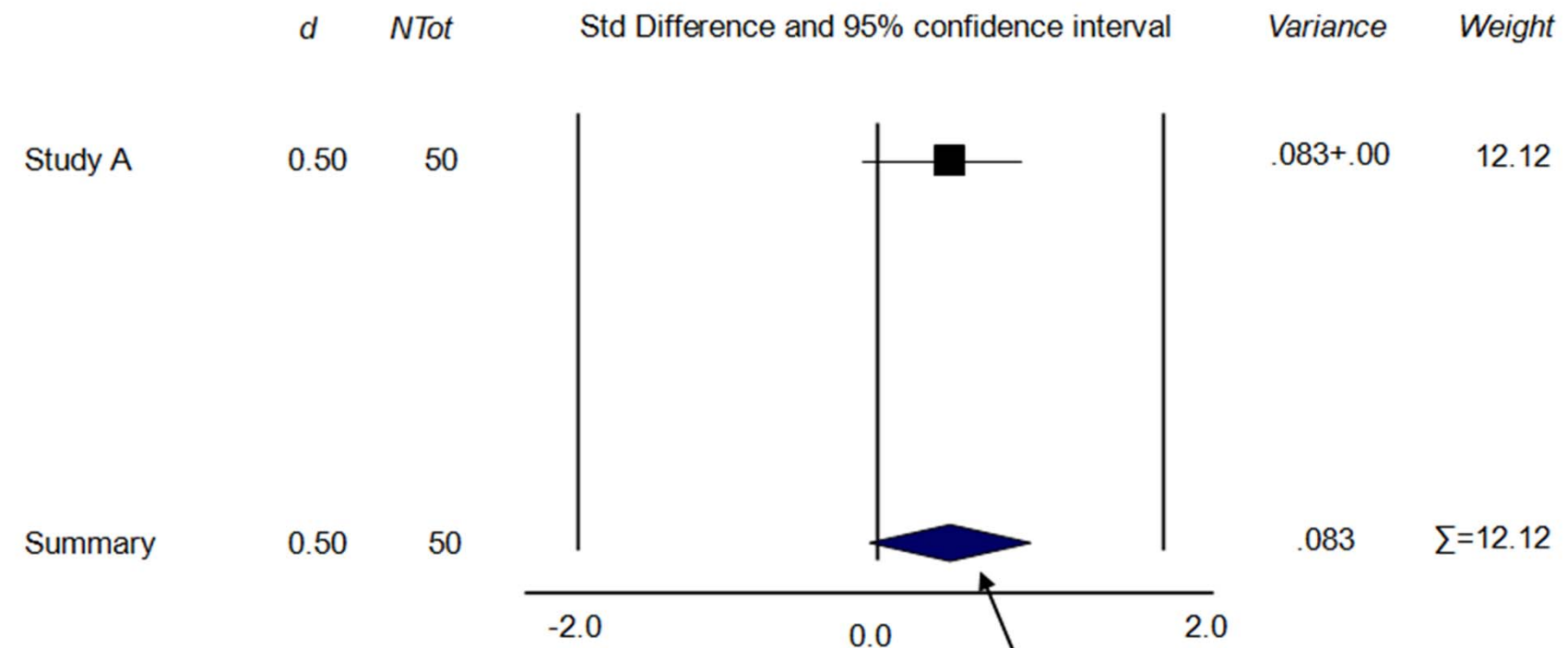


Aptitude score at a sample of colleges



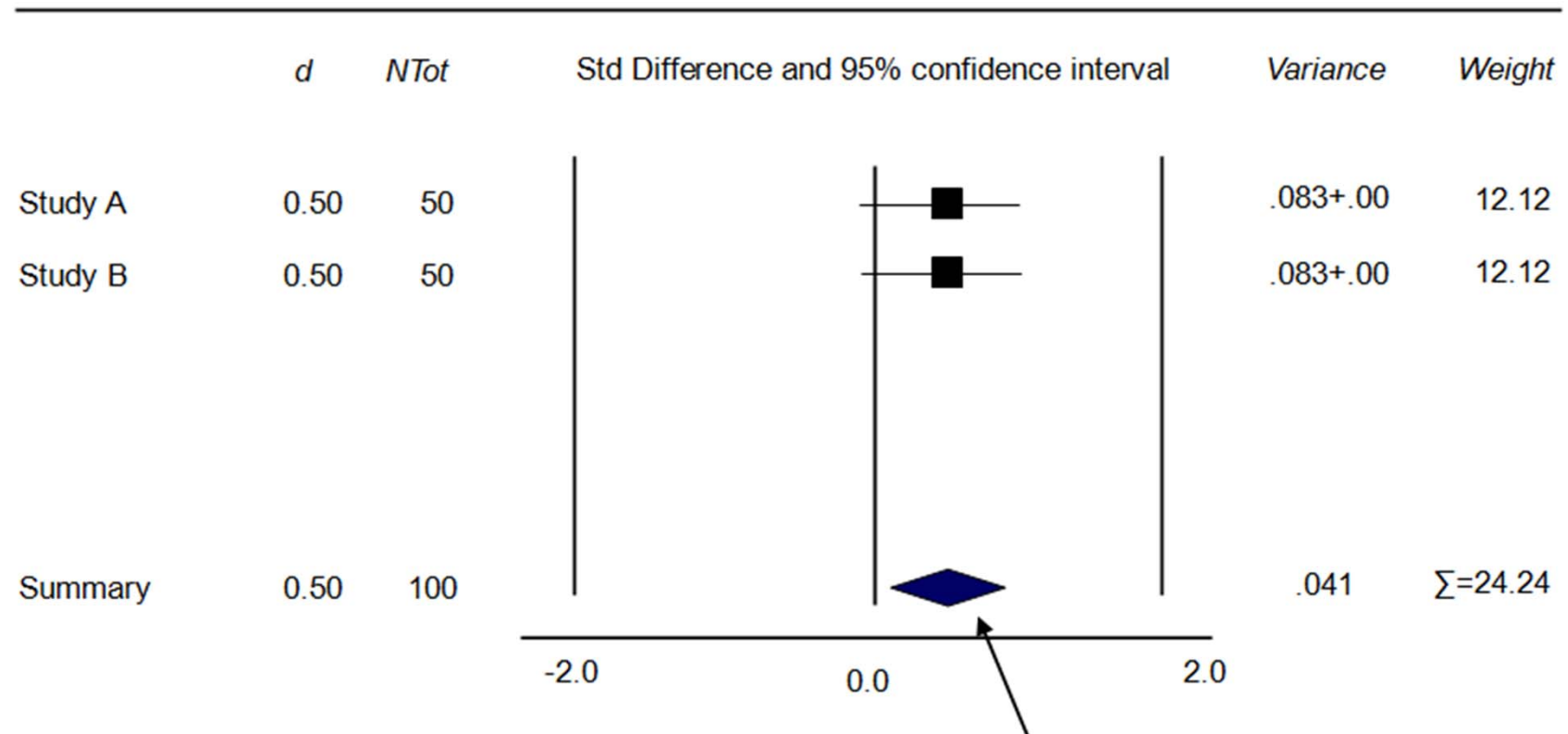
Impact on substantive utility

Meta-analysis with consistent effects $k = 1$



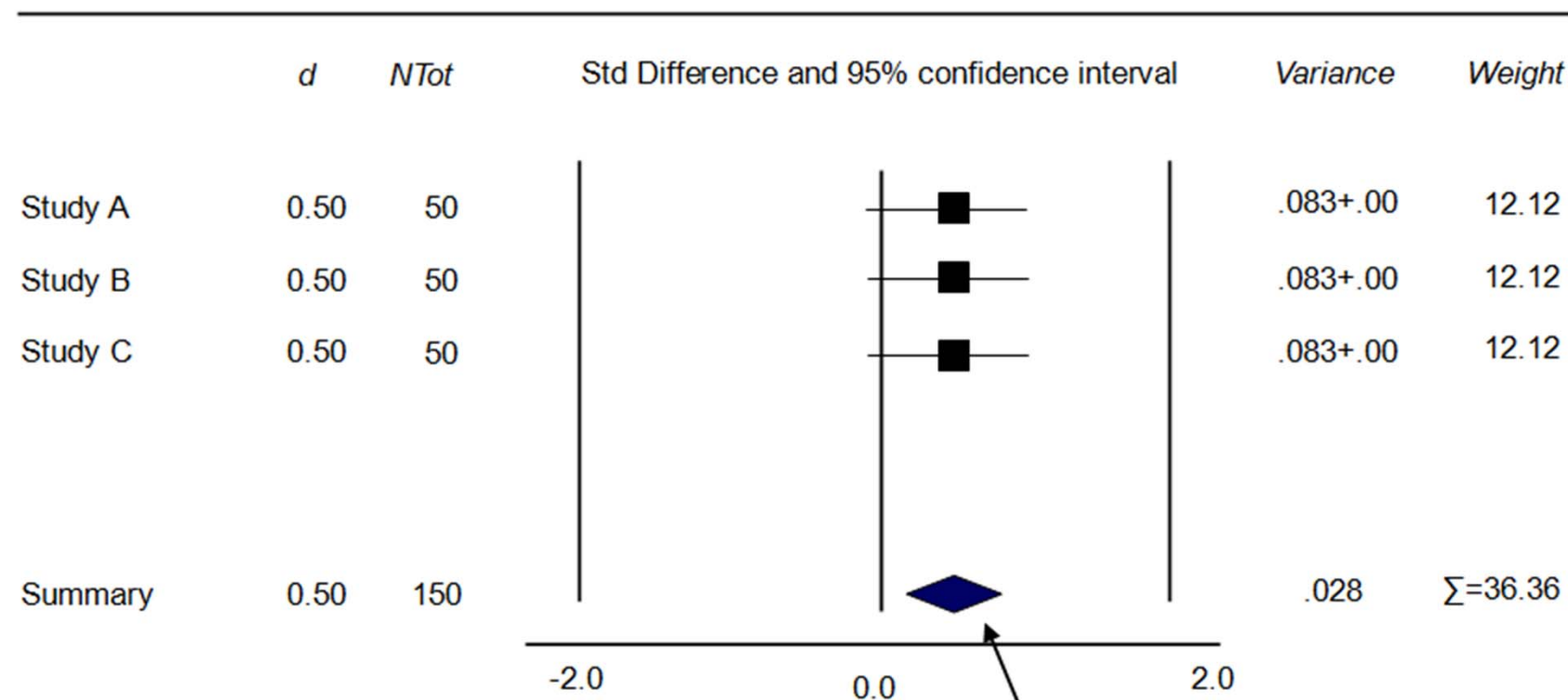
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{6.06}{12.12} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{12.12} = 0.083 \quad SE = \sqrt{0.083} = 0.287$$

Meta-analysis with consistent effects $k = 2$



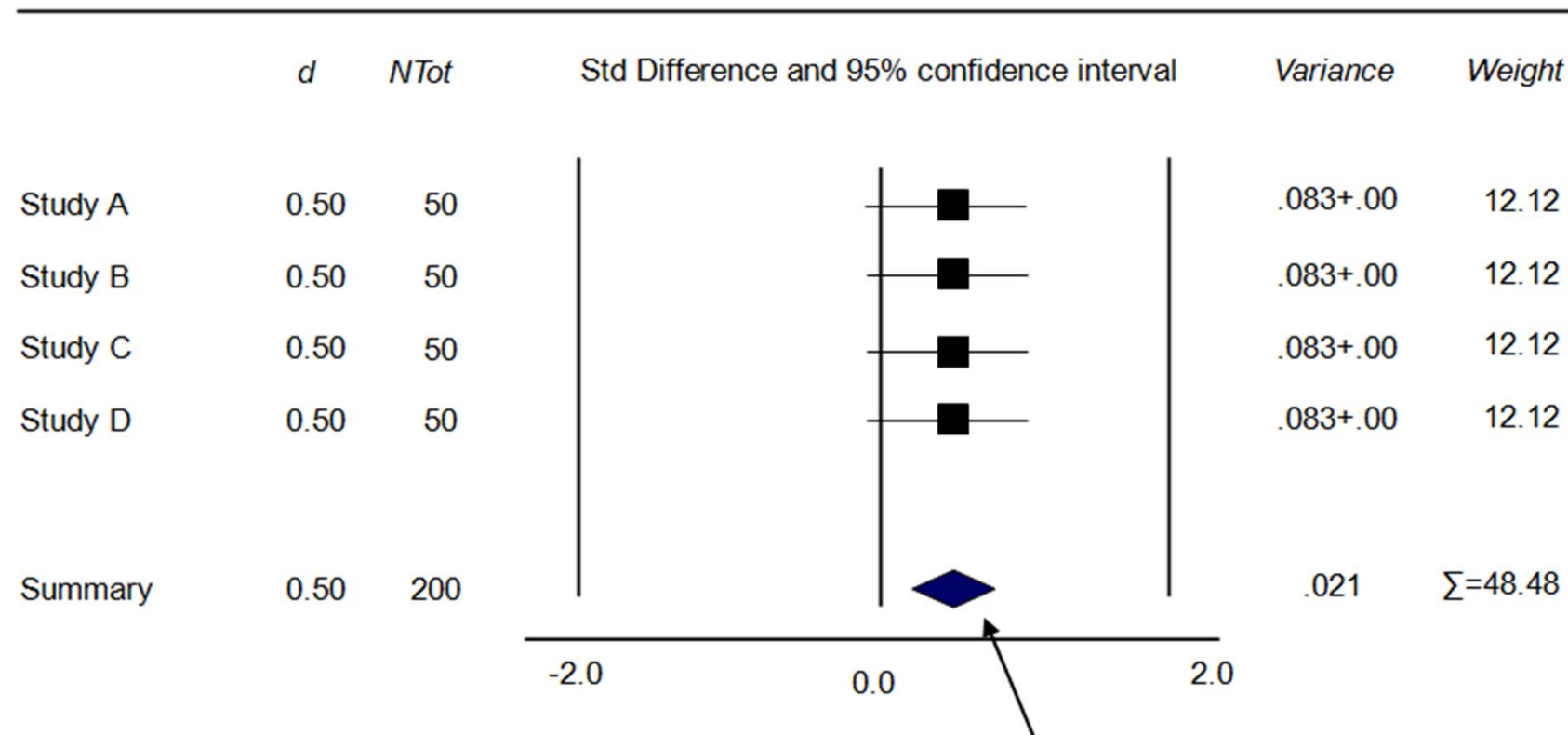
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{12.12}{24.24} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{24.24} = 0.041 \quad SE = \sqrt{0.041} = 0.203$$

Meta-analysis with consistent effects $k = 3$



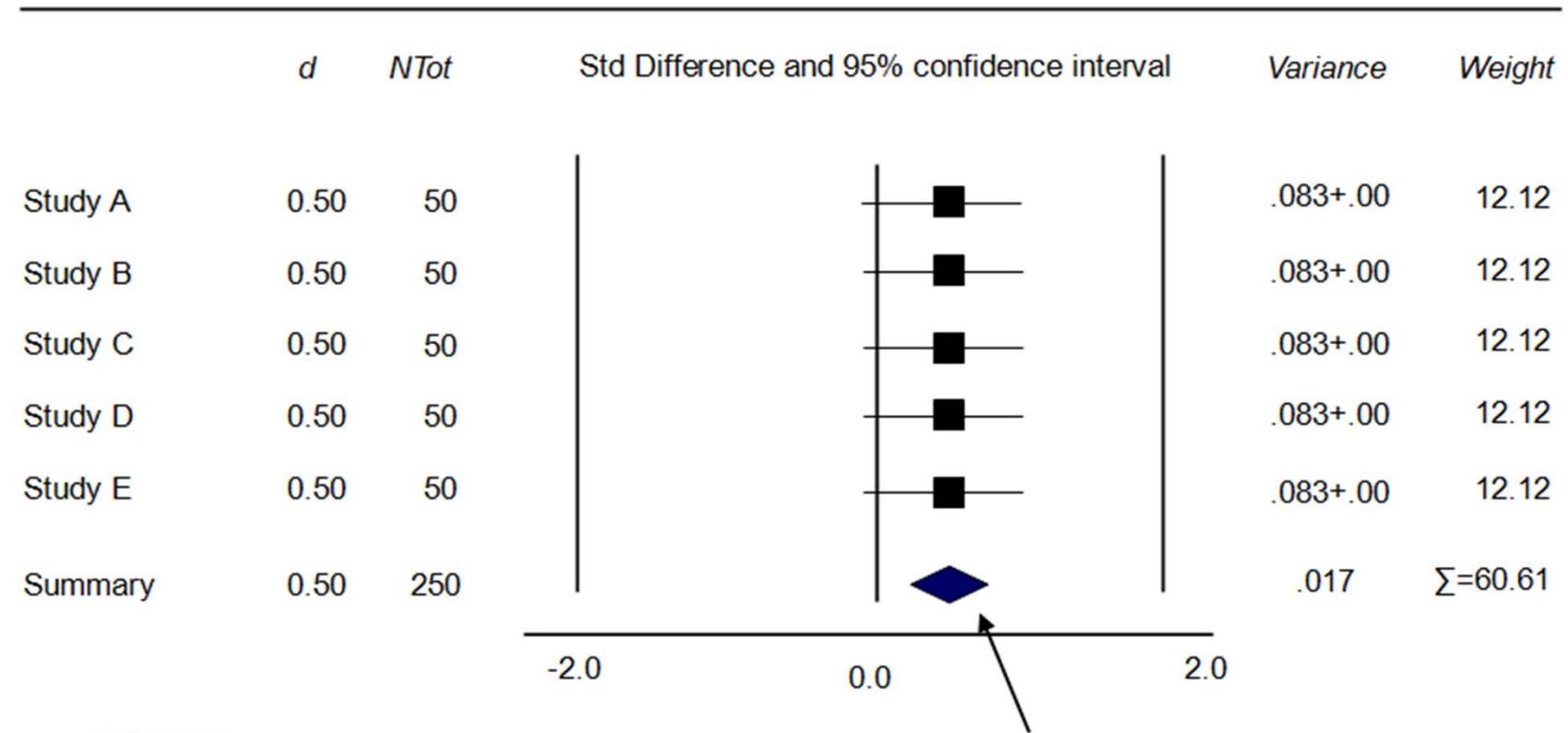
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{18.18}{36.36} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{36.36} = 0.028 \quad SE = \sqrt{0.028} = 0.166$$

Meta-analysis with consistent effects $k = 4$



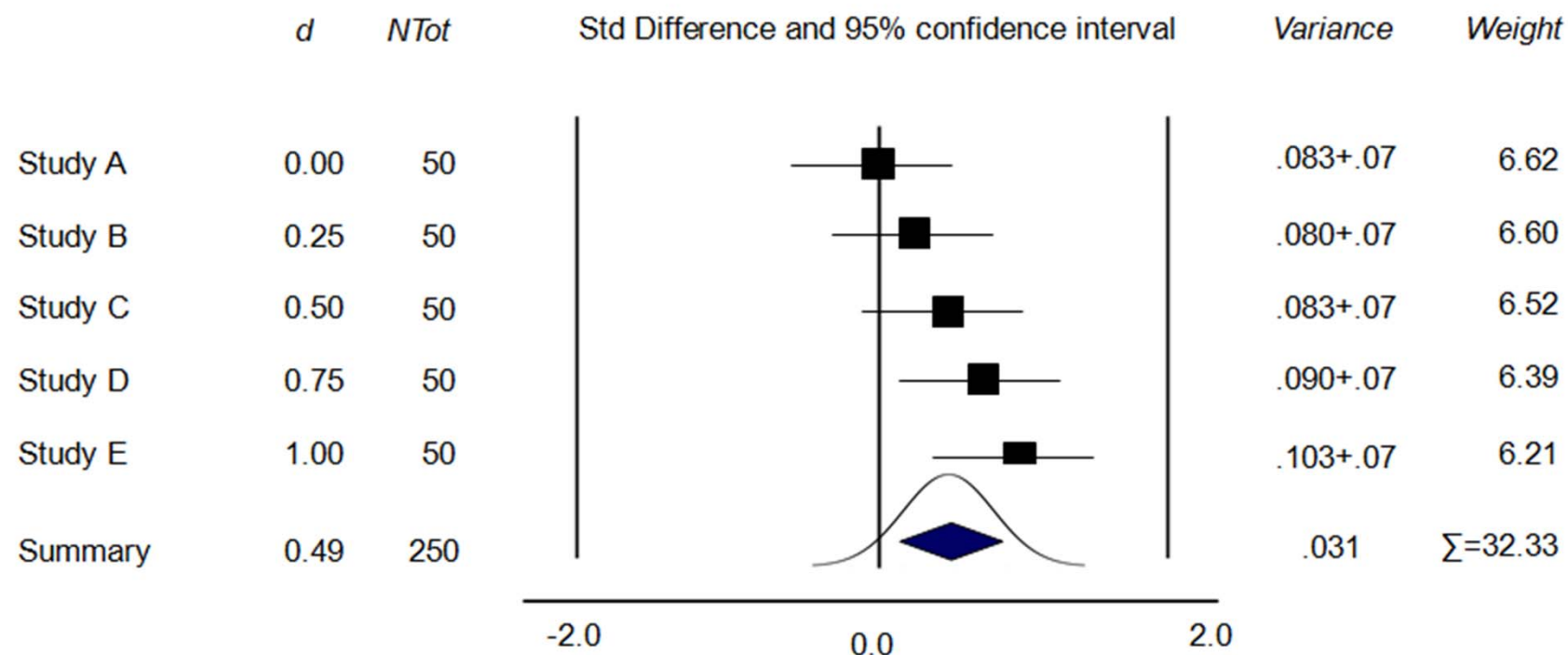
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{24.24}{48.48} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{48.48} = 0.021 \quad SE = \sqrt{0.021} = 0.144$$

Meta-analysis with consistent effects $k = 5$



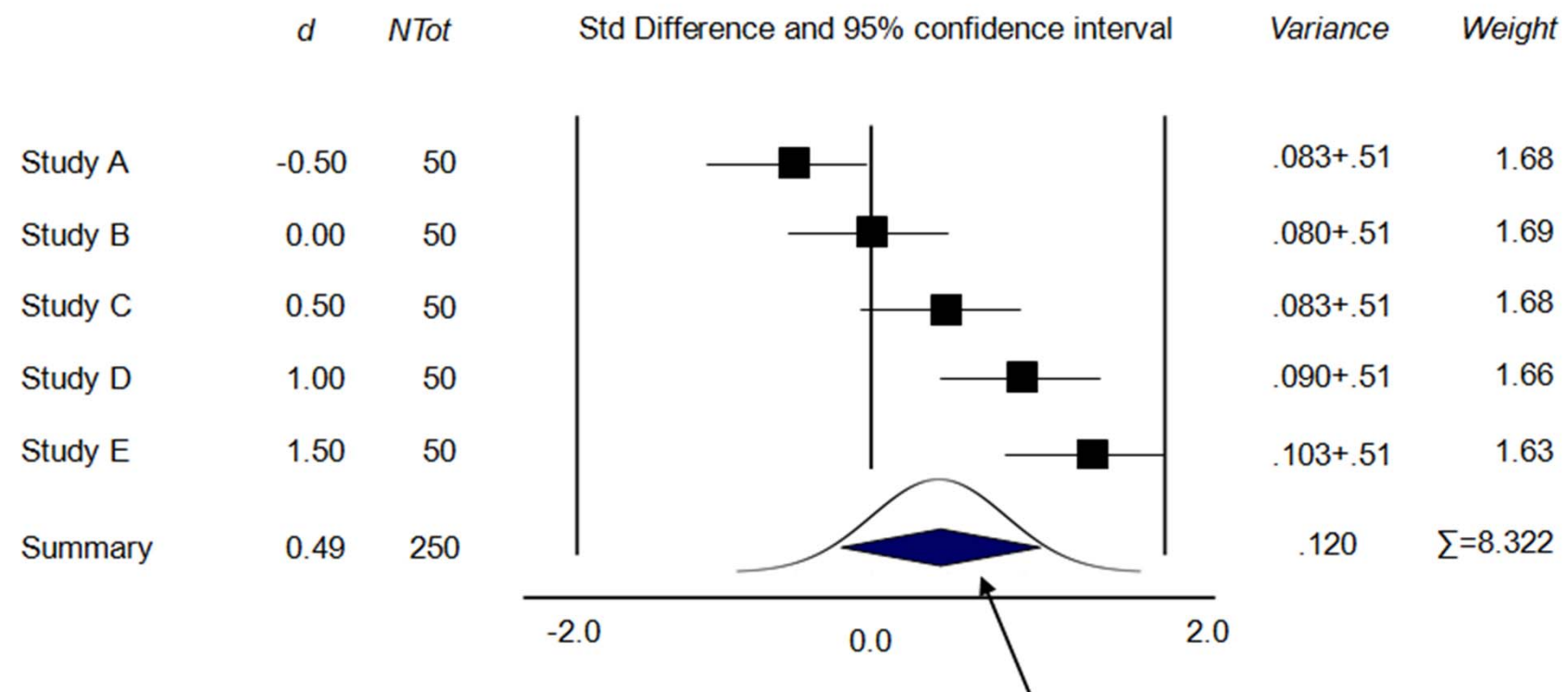
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{30.30}{60.60} = 0.50 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{60.61} = 0.017 \quad SE = \sqrt{0.017} = 0.128$$

Meta-analysis with heterogeneous effects $k = 5$



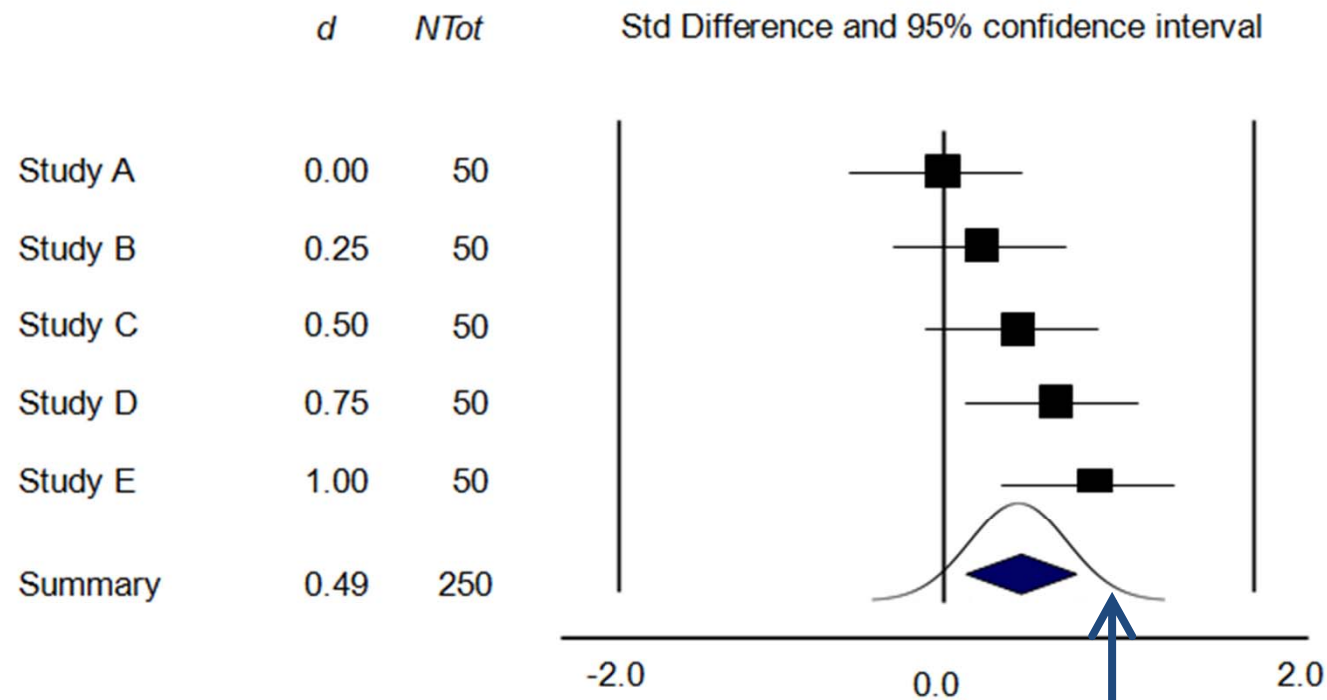
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{15.91}{32.32} = 0.49 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{32.32} = 0.031 \quad SE = \sqrt{0.031} = 0.176$$

Meta-analysis with heterogeneous effects $k = 5$



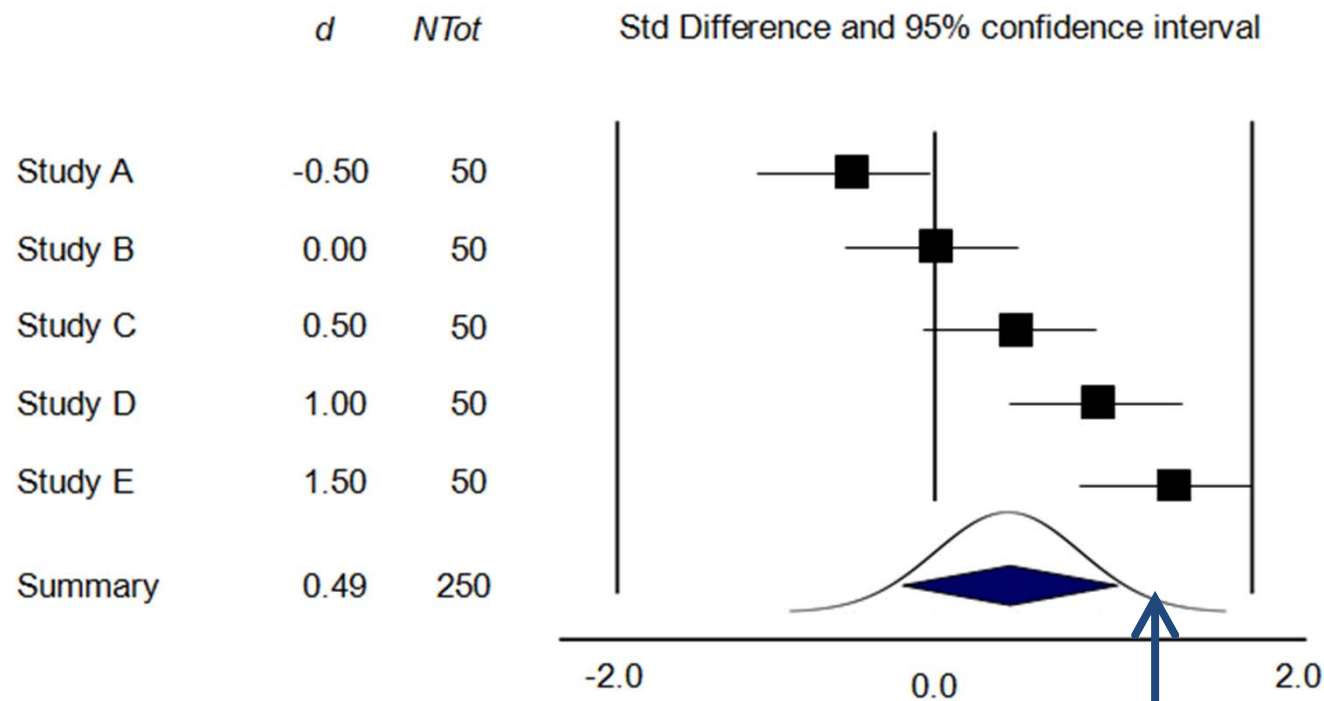
$$M = \frac{\sum W_i Y_i}{\sum W_i} = \frac{4.09}{8.32} = 0.49 \quad V_M = \frac{1}{\sum W_i} = \frac{1}{8.32} = 0.120 \quad SE = \sqrt{0.120} = 0.347$$

Meta-analysis with heterogeneous effects $k = 5$



Dispersion of the
individual effects

Meta-analysis with heterogeneous effects $k = 5$



Dispersion of the
individual effects

That's *why* we need to quantify

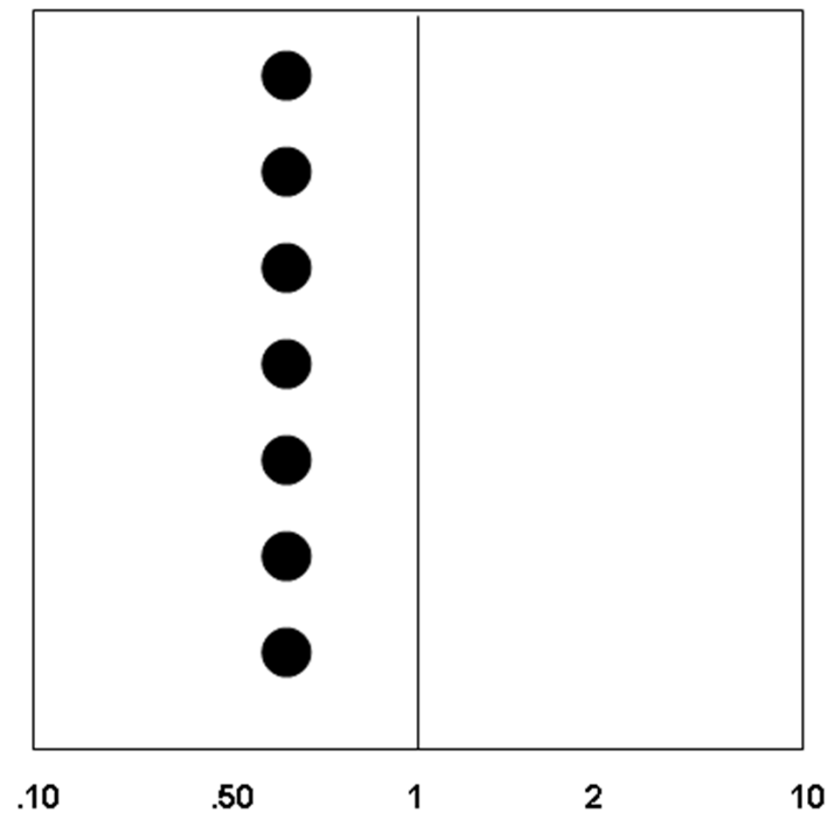
How do we quantify heterogeneity?

Two-step process

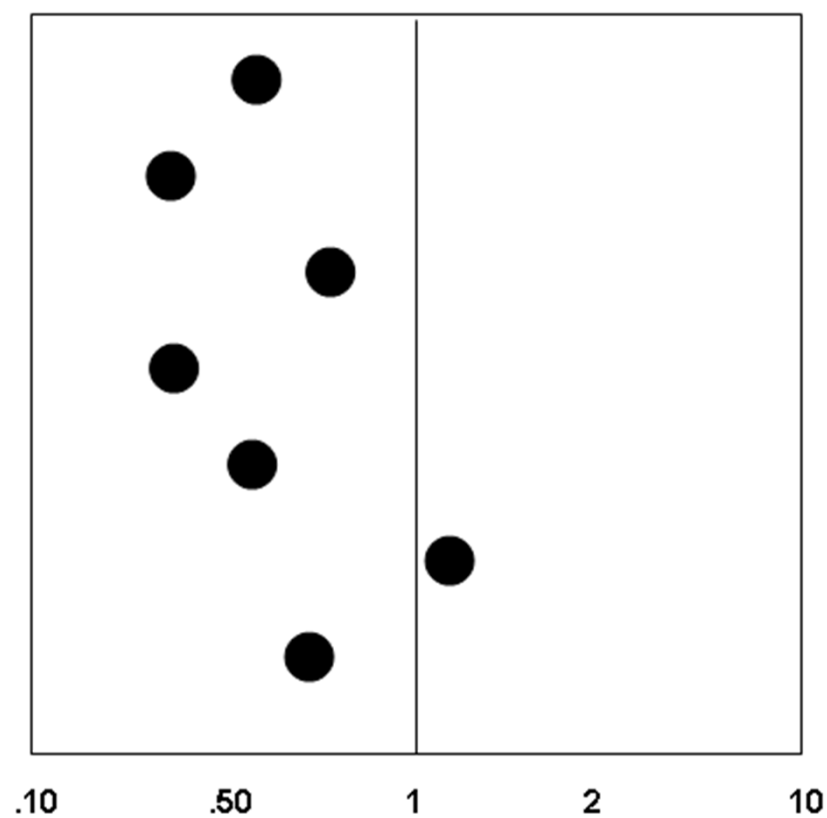
- Isolate the real dispersion
- Translate this into useful indices

Part 1 – Isolate the real dispersion

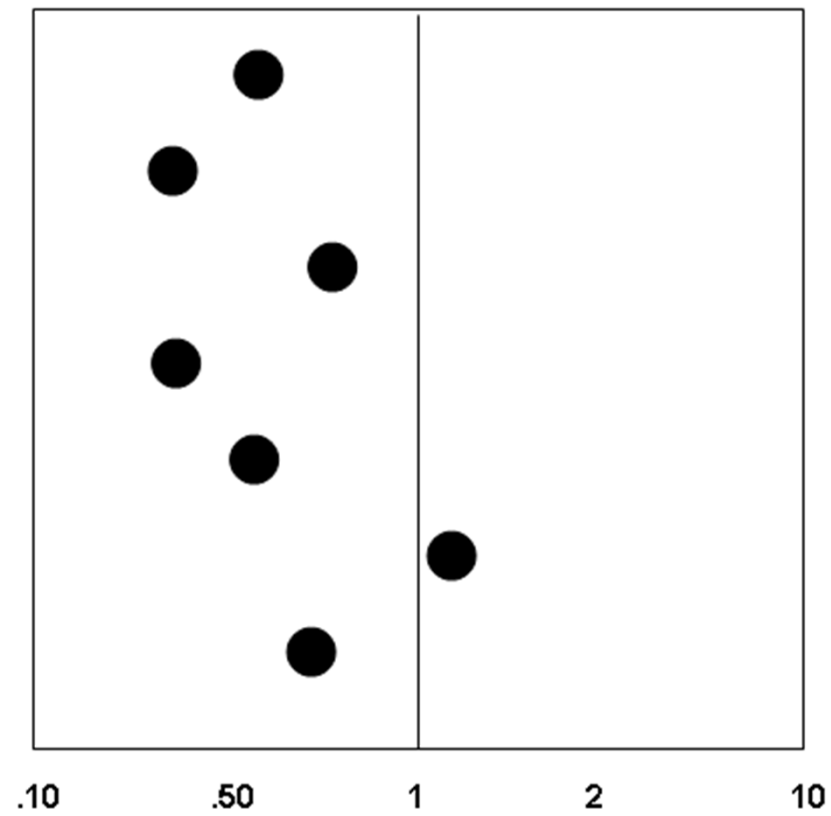
What we'd like to see if the true effect is the same in all studies



What we might see if the true effect is the same in all studies

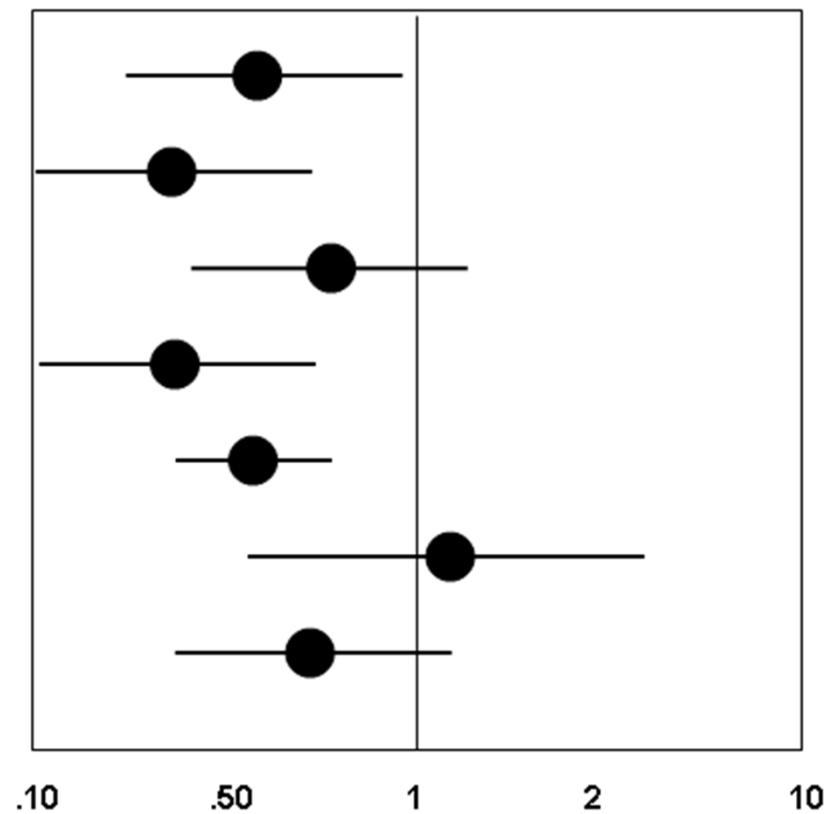


Is there real dispersion?



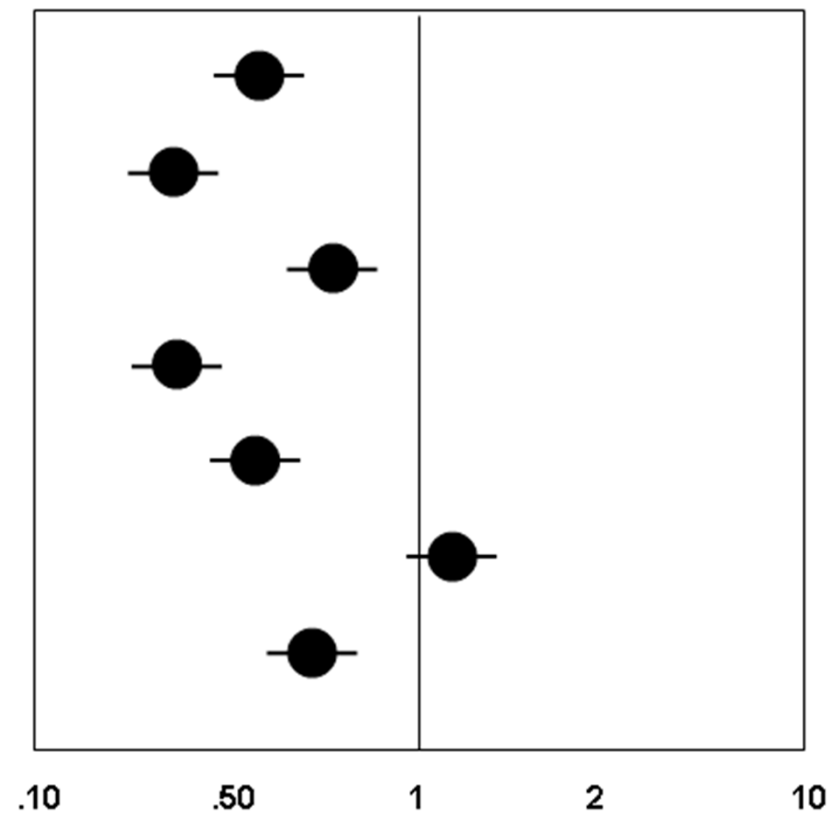
Split, Croatia June 2014

It depends on the precision



Split, Croatia June 2014

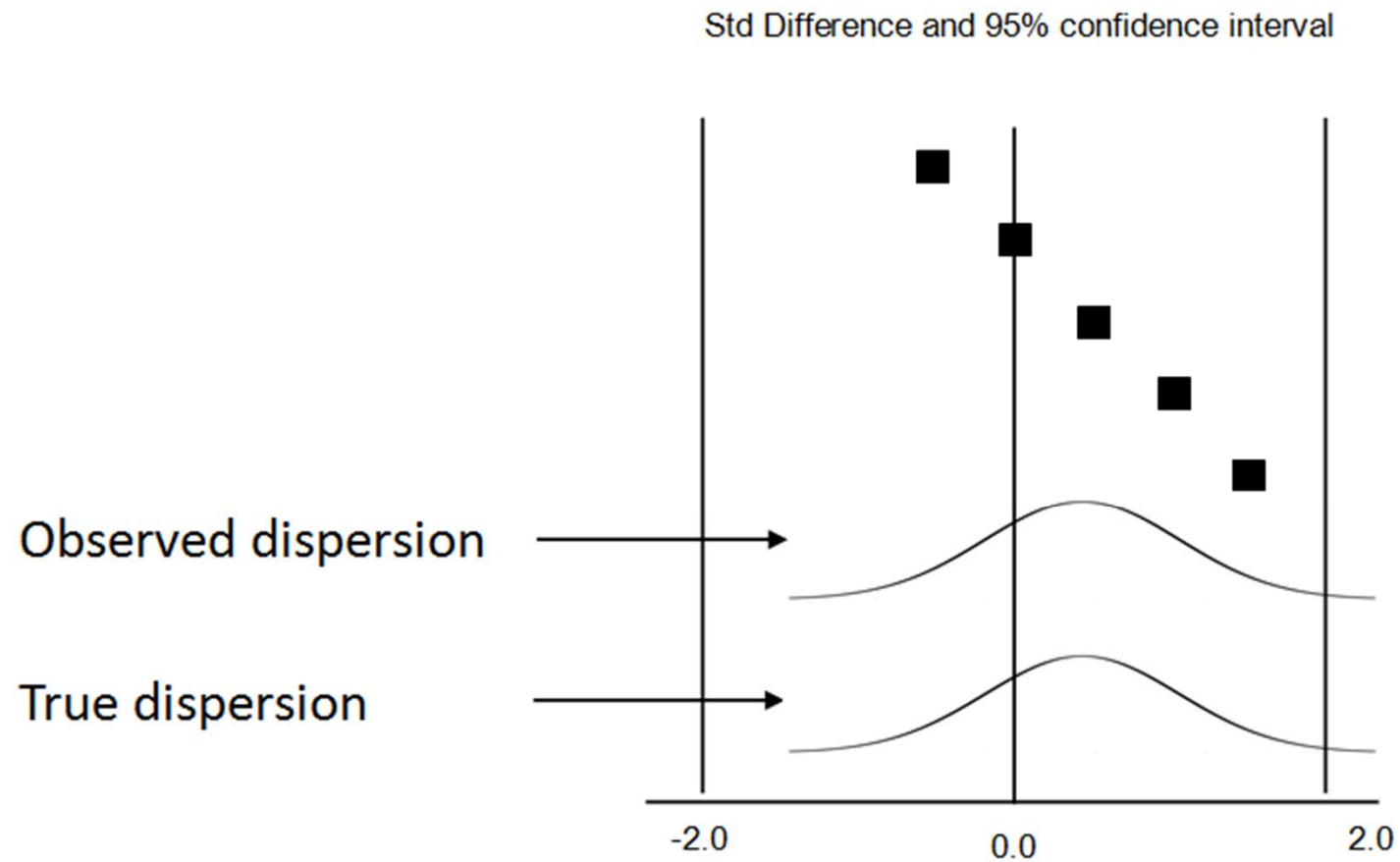
It depends on the precision



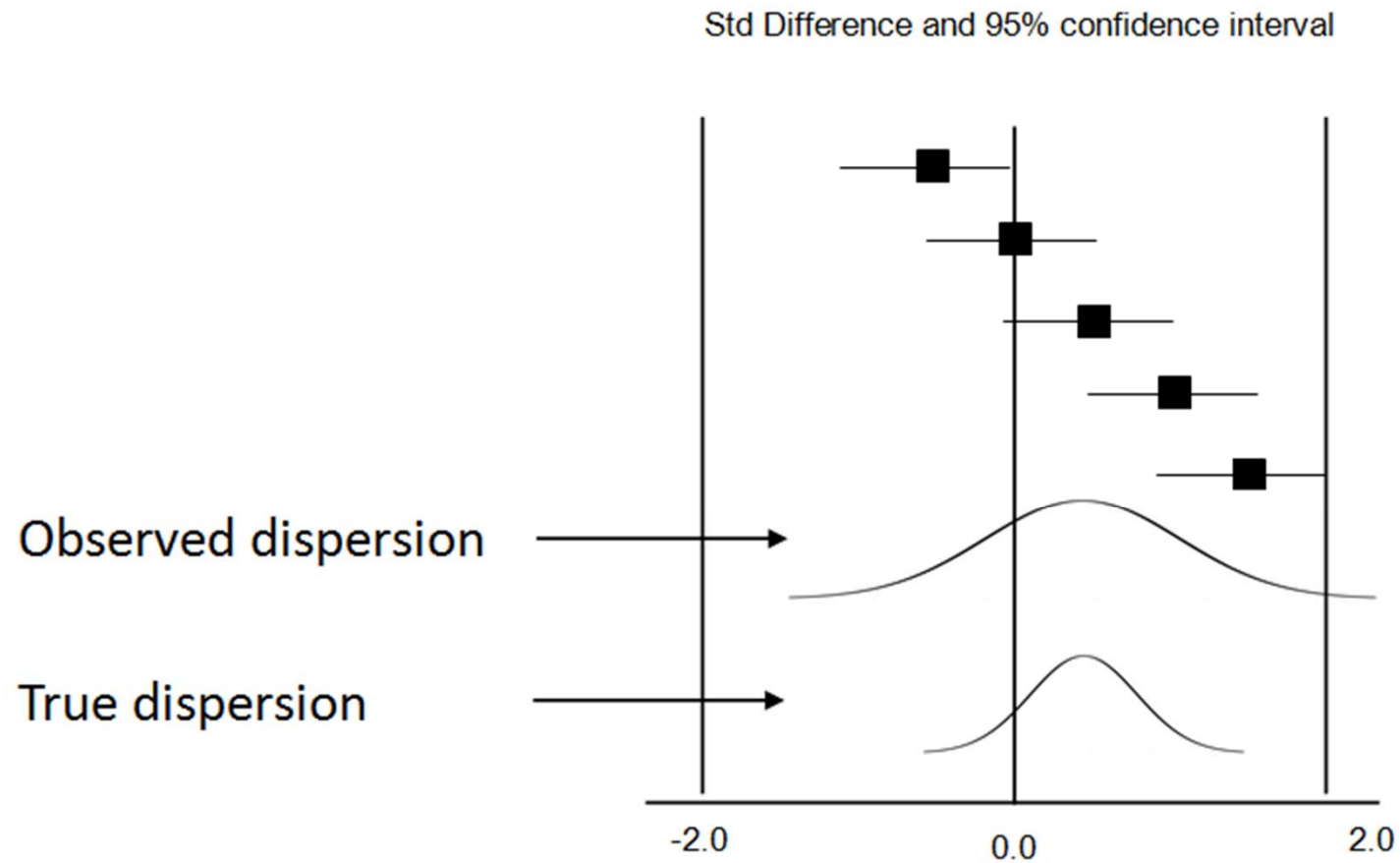
Key point

- We can easily compute variance of *observed* effects
- But this is due partly to real differences in effects and partly to sampling error within studies
- We need to isolate the between-studies variance

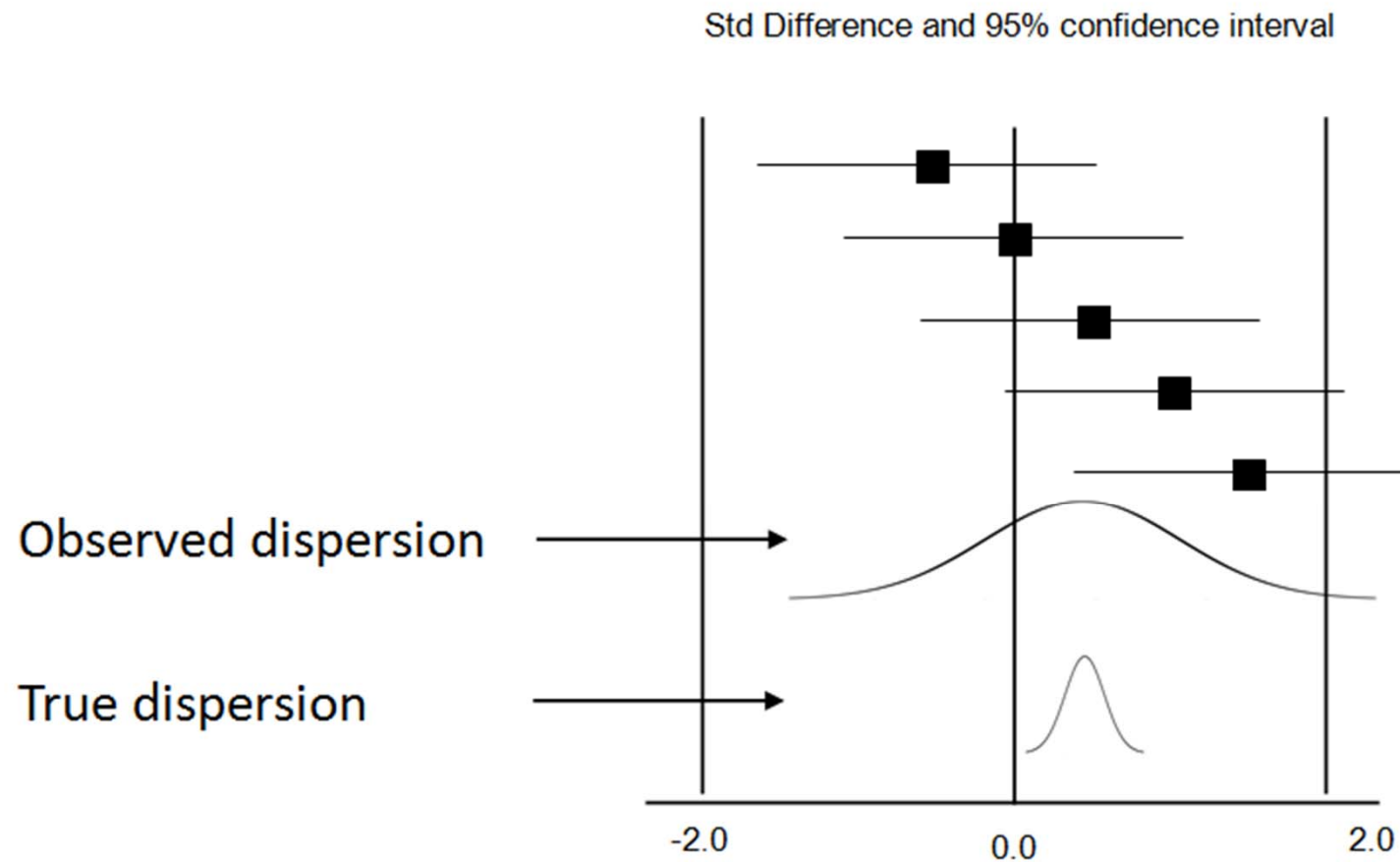
Going from observed to true heterogeneity



Going from observed to true heterogeneity



Going from observed to true heterogeneity



To assess heterogeneity

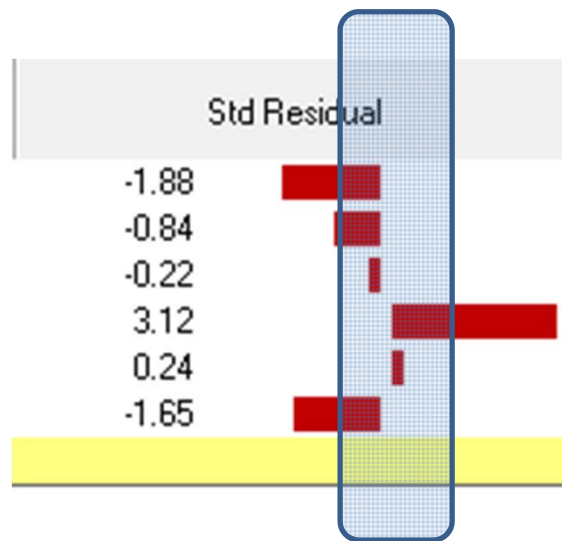
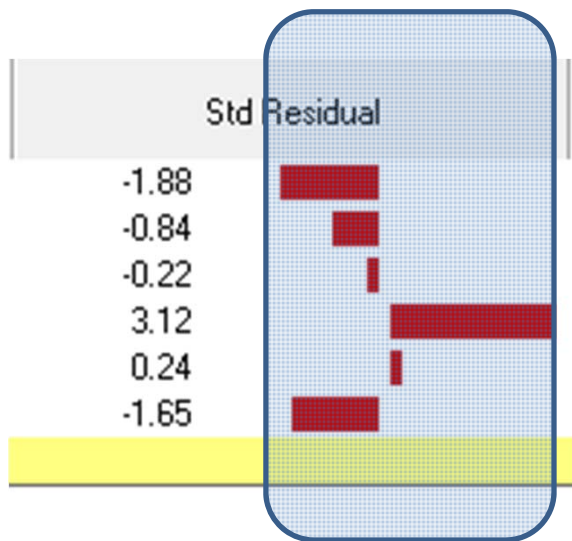
- Compute observed variance
- Estimate how much variance would be expected if true effect is identical in all studies
- Observed minus expected is estimate of true variance

Isolating the real dispersion

Q

df

$Q-df$



Indices related to heterogeneity

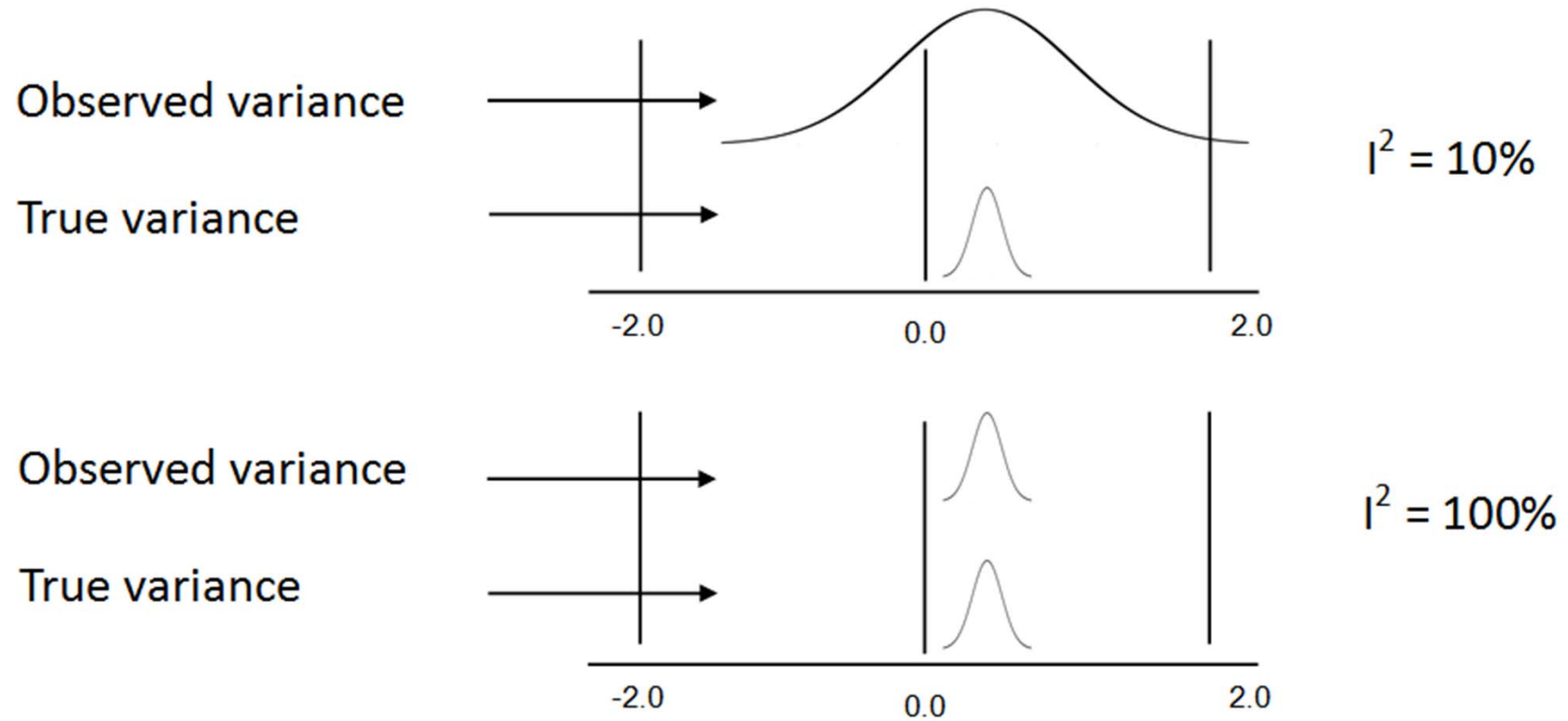
$Q-df$	Basis for all indices
p	Test of null
T	Standard deviation of true effects
T^2	Variance of true effects
I^2	Proportion of true/total variance

P -value

- Can we conclude that there is some variance in true effects
- Depends on amount of excess variance *and* the amount of evidence

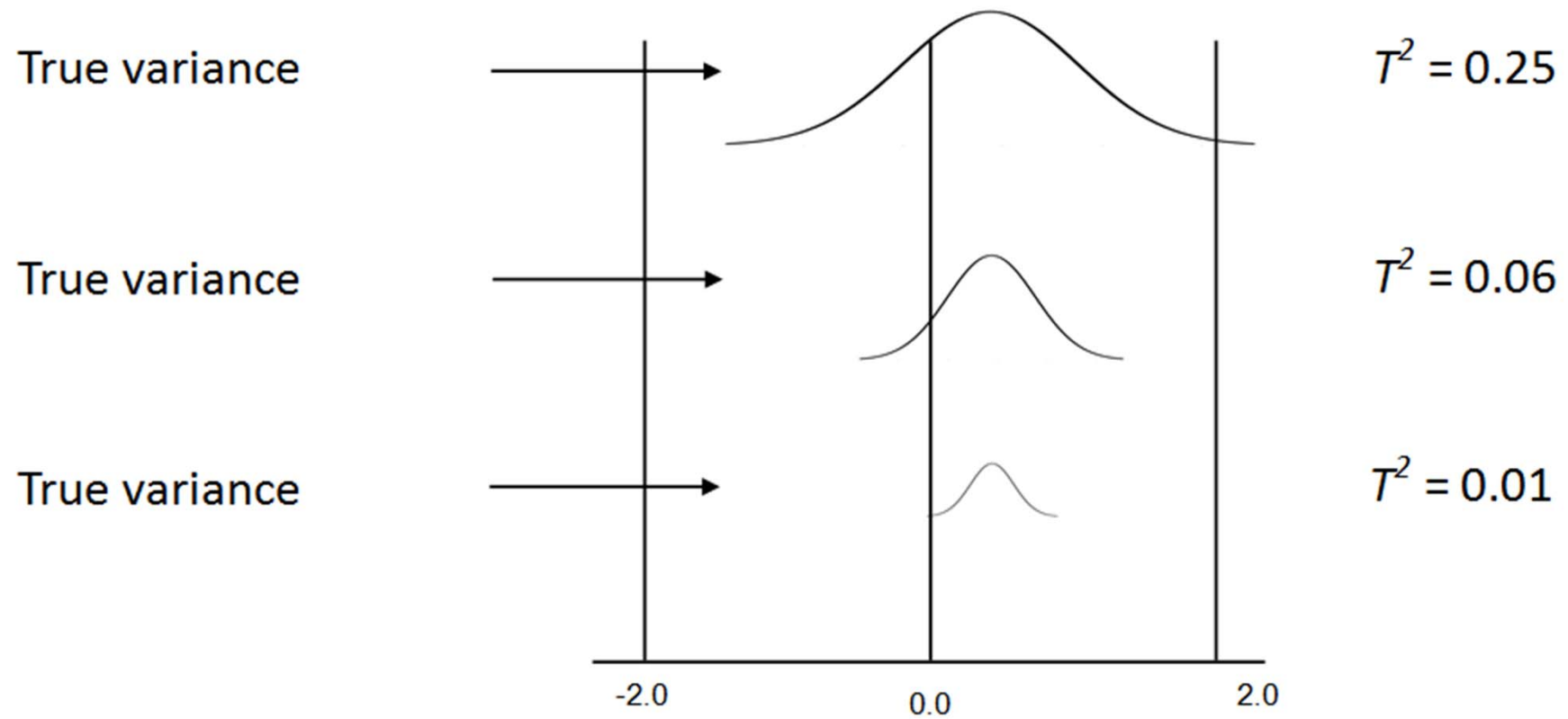
What proportion of variance is real? I^2

Std Difference and 95% confidence interval

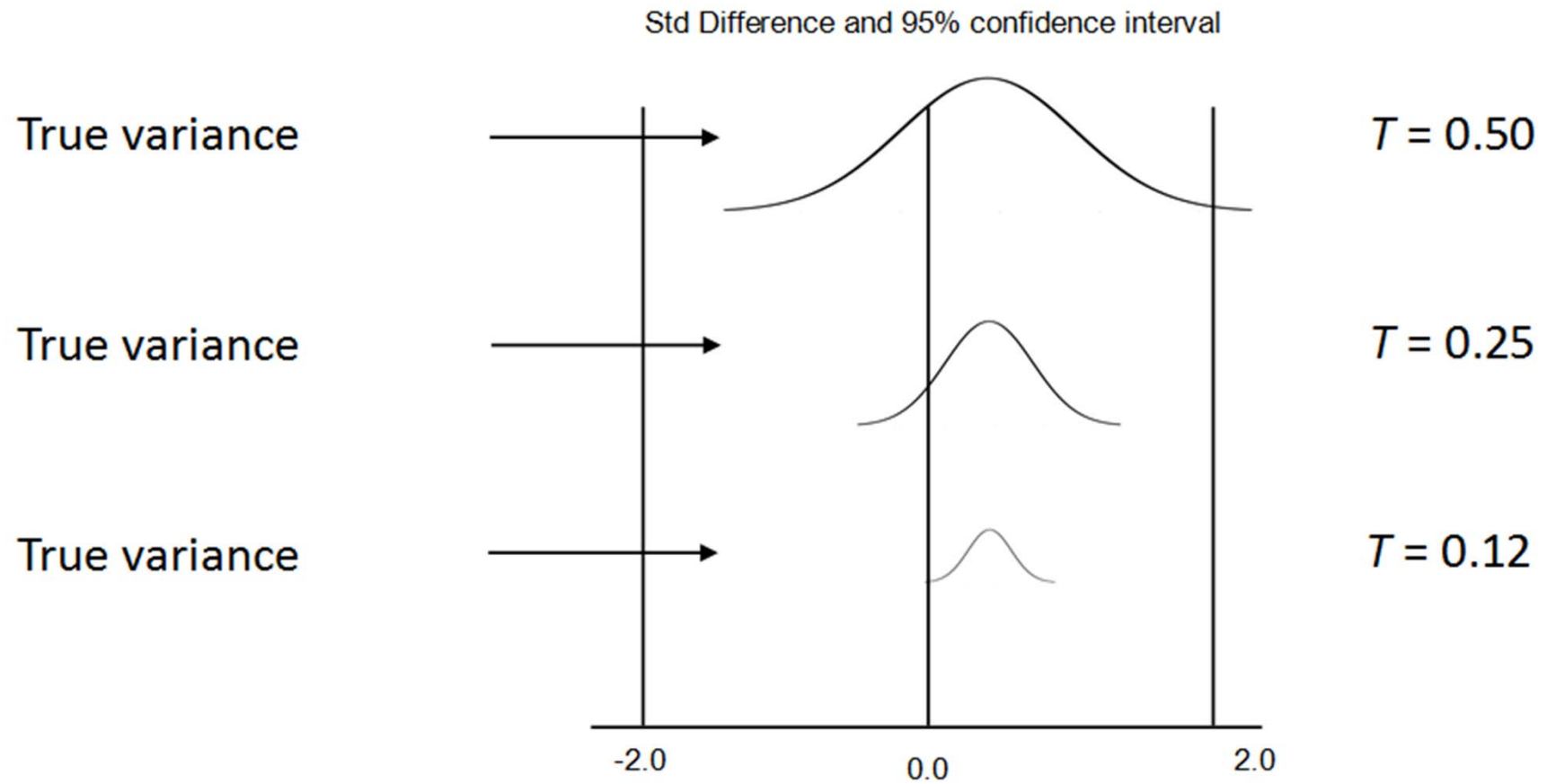


How much real dispersion (squared)? τ^2

Std Difference and 95% confidence interval



How much real dispersion (non-squared)? T



Statistics apply to both models

Comprehensive meta analysis - [Analysis]

File Edit Format View Computational options Analyses Help

← Data entry ↗ Next table High resolution plot Select by ... Effect measure: Hedges's g

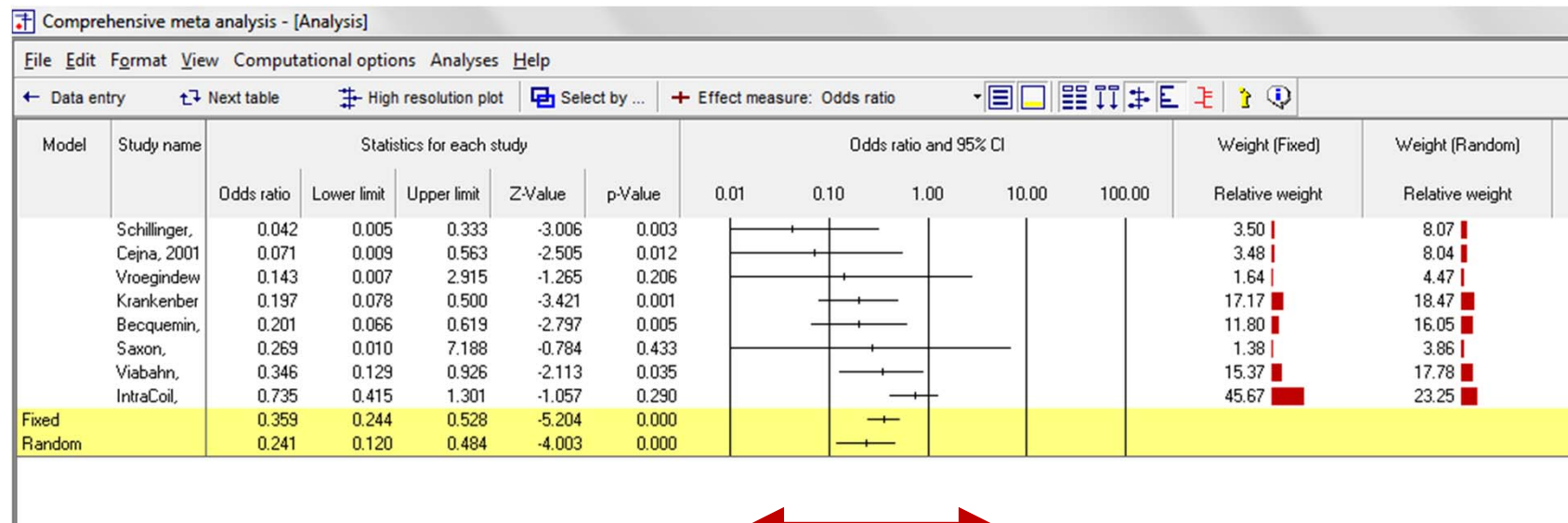
Model	Effect size and 95% confidence interval						Test of null (2-Tail)		Heterogeneity				Tau-squared			
Model	Number Studies	Point estimate	Standard error	Variance	Lower limit	Upper limit	Z-value	P-value	Q-value	df (Q)	P-value	I-squared	Tau Squared	Standard Error	Variance	Tau
Fixed	6	0.414	0.064	0.004	0.289	0.540	6.474	0.000	12.003	5	0.035	58.345	0.037	0.042	0.002	0.193
Random	6	0.358	0.105	0.011	0.152	0.565	3.404	0.001								

Caution !

- I^2 is *NOT* a measure of absolute heterogeneity
- I^2 tells us what proportion of the observed dispersion reflects differences in true scores rather than random sampling error

I^2

What proportion of the observed variance is real?



Fixed-effect vs. Random-effects

Fixed-effect vs. Random-effects

- Fixed-effect
 - Sampling takes place at one level only
 - Any between-study variance will be ignored when assigning weights
- Random-effects
 - Sampling takes place at two levels
 - Any between-study variance will be used when assigning weights

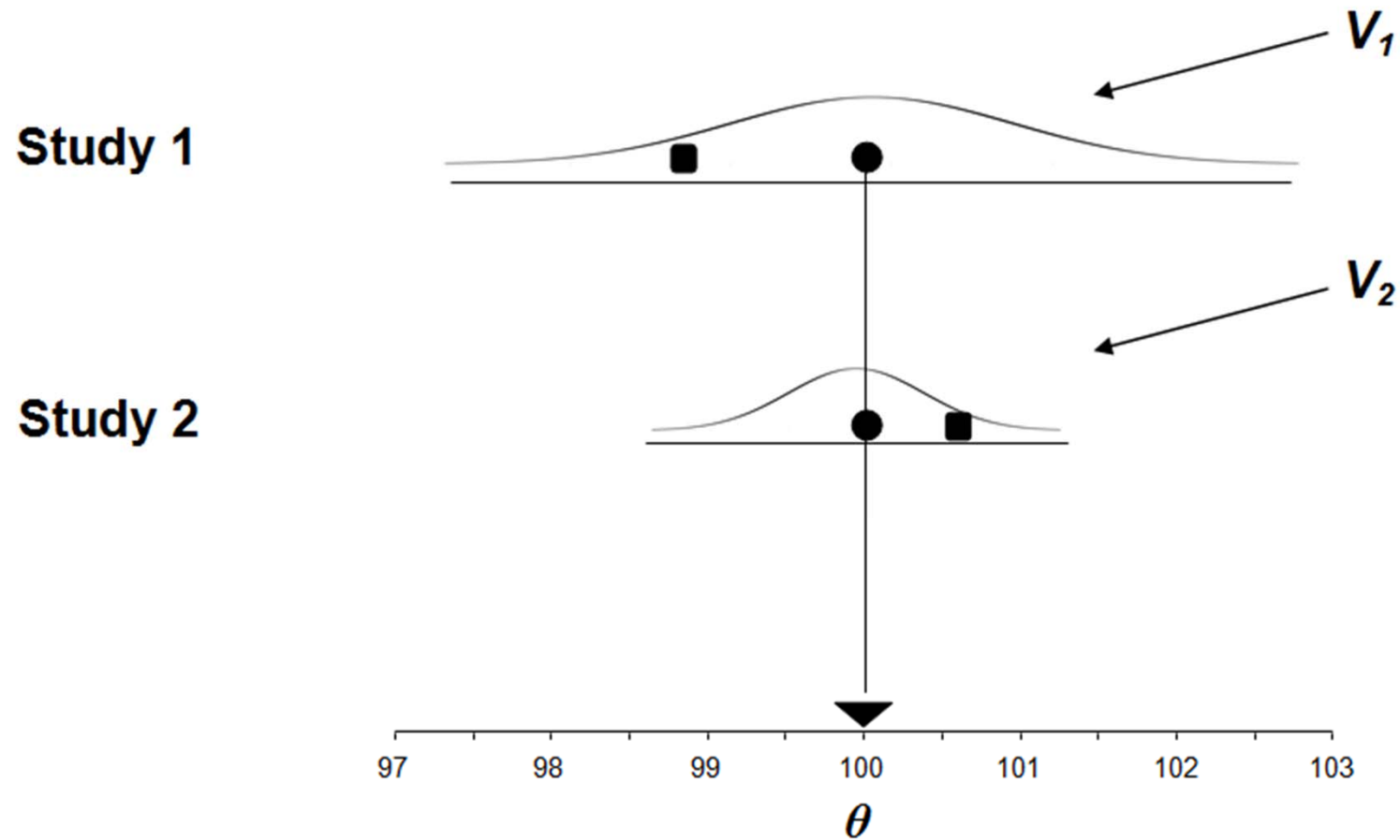
Fixed effect

- When there is reason to believe that all the studies are functionally identical
- When our goal is to compute the common effect size, for the studies in the analysis
- Example of drug company has run five studies to assess the effect of a drug.

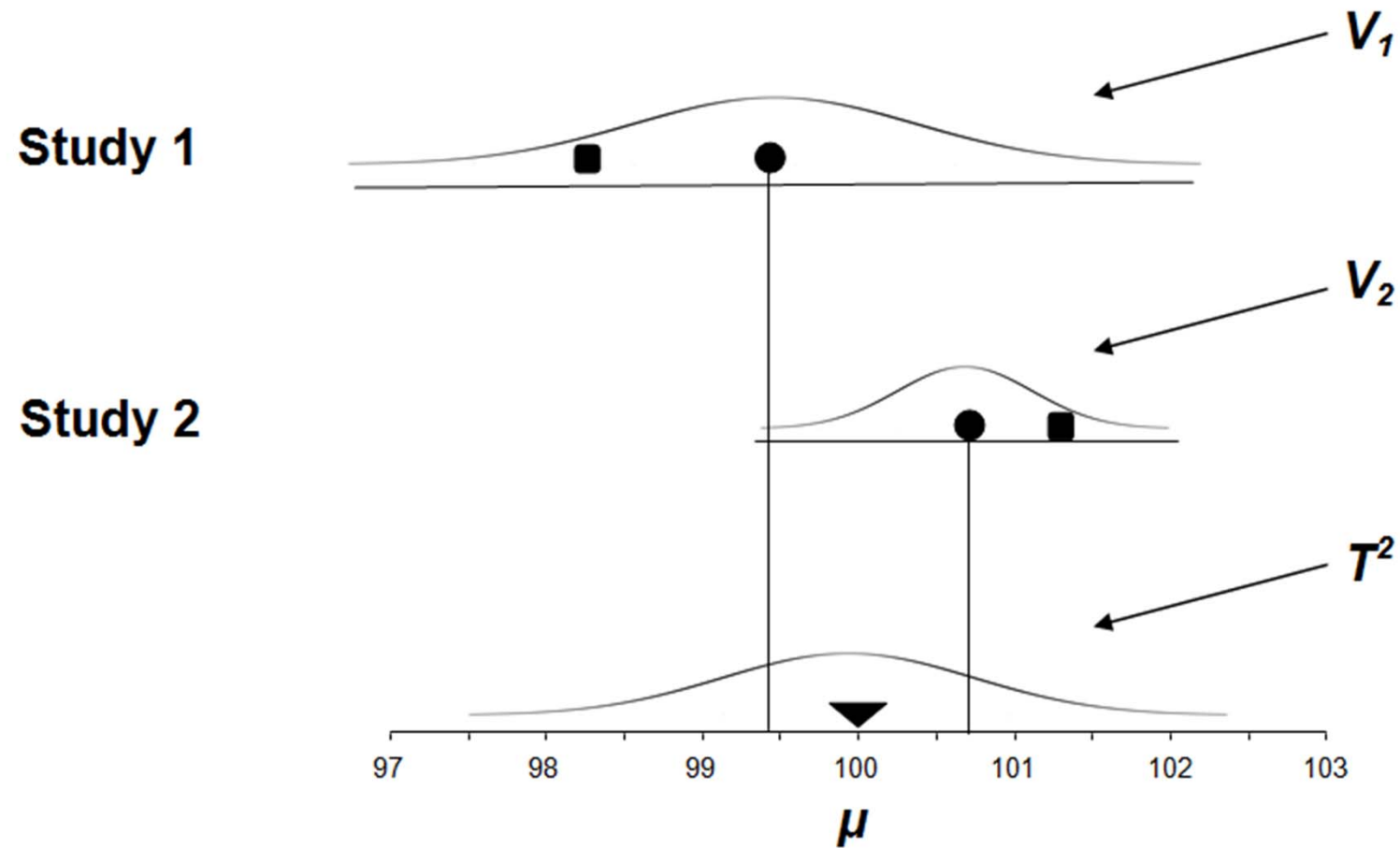
Random effects

- When not likely that all the studies are functionally equivalent.
- When the goal of this analysis is to generalize to a range of populations.
- Example of studies culled from publications

Sampling error under fixed-effect model



Sampling error under random-effects model



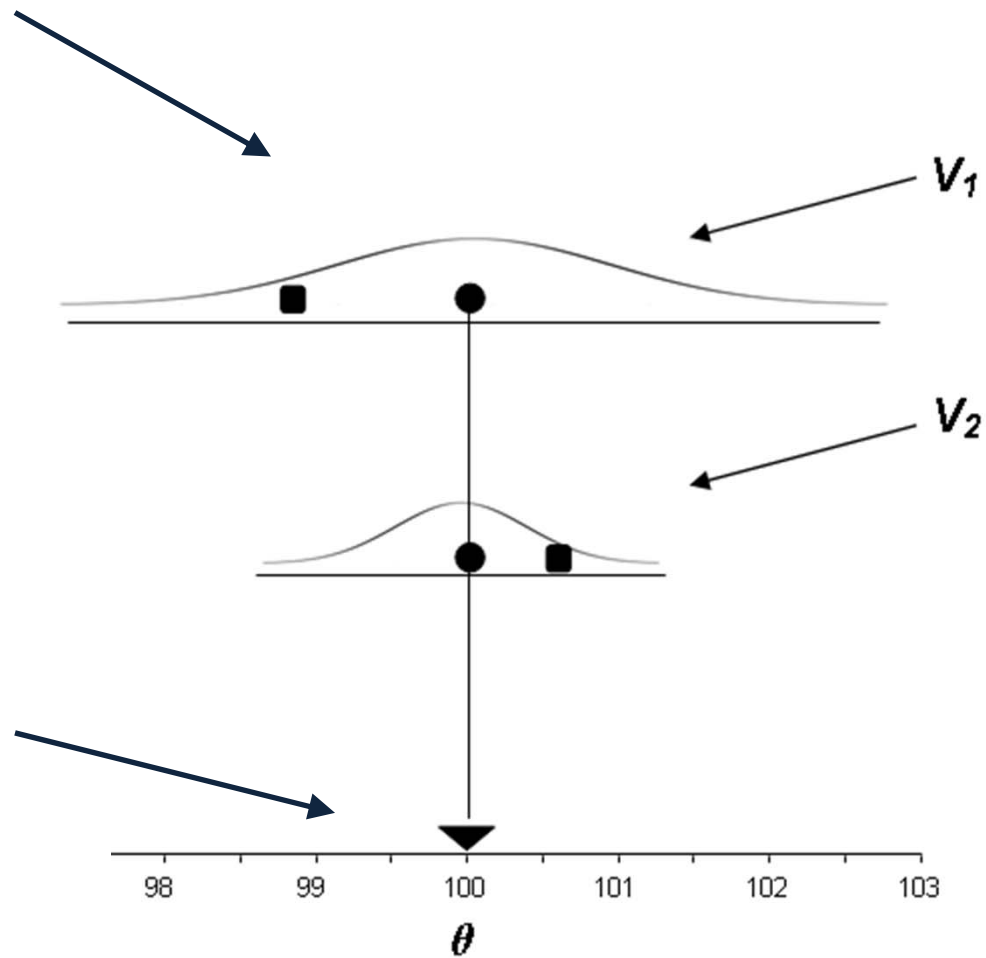
Definition of combined effect

- Fixed effect model
 - There is one true effect
 - Summary effect is estimate of this value
- Random effects model
 - There is a distribution of effects
 - Summary effect is mean of distribution

How weights are assigned

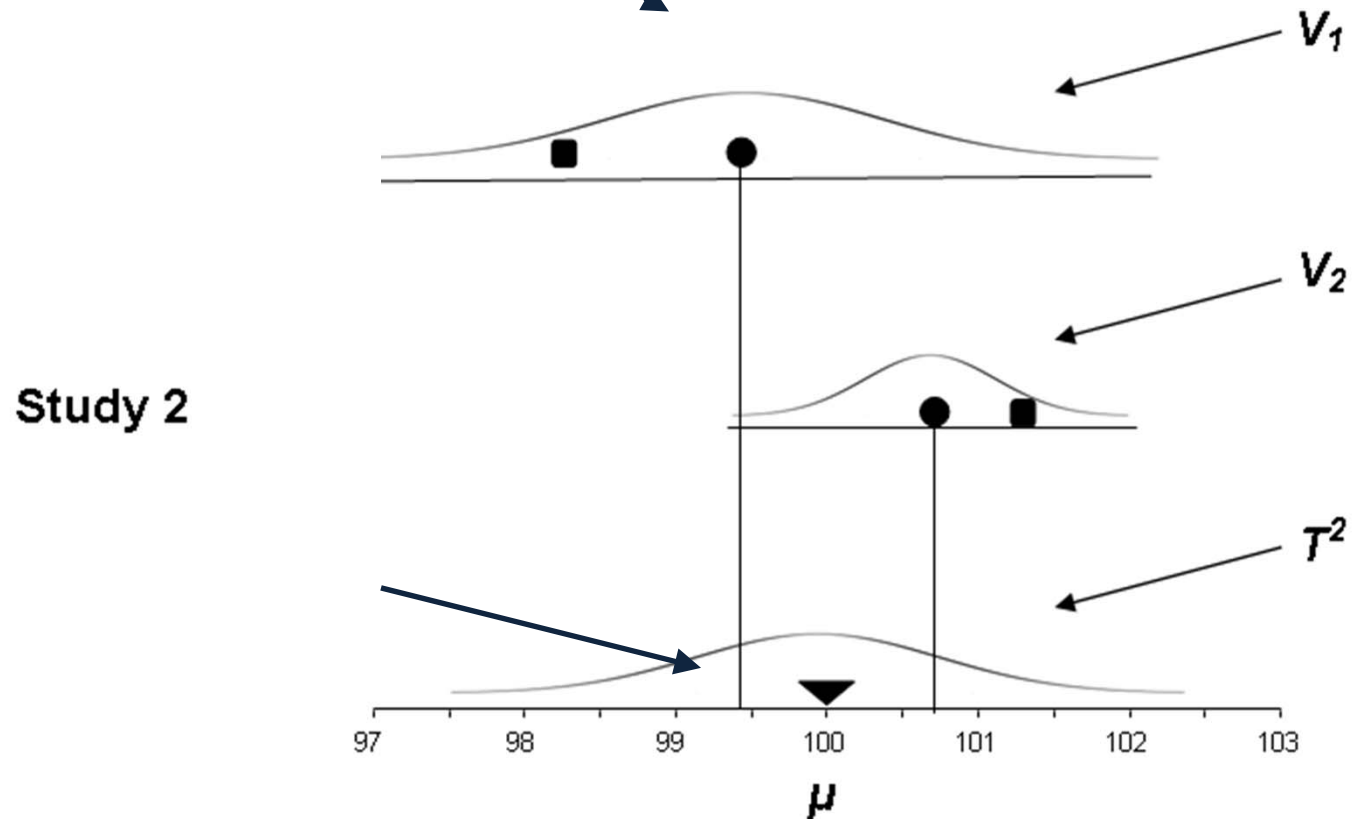
Fixed effect model

$$W = 1 / (V_1)$$



Random-effects model

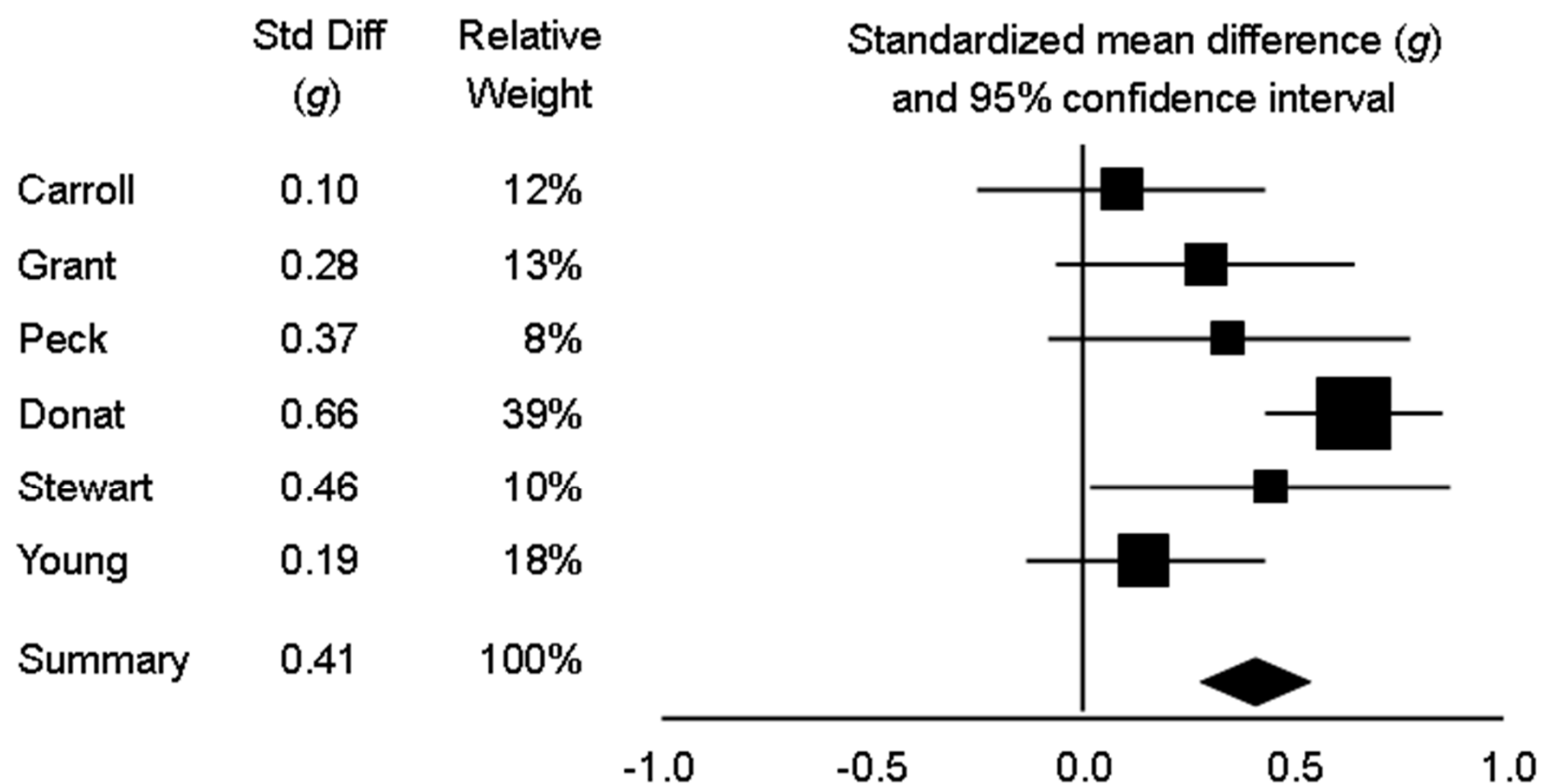
$$W = 1 / (V_1 + T^2)$$



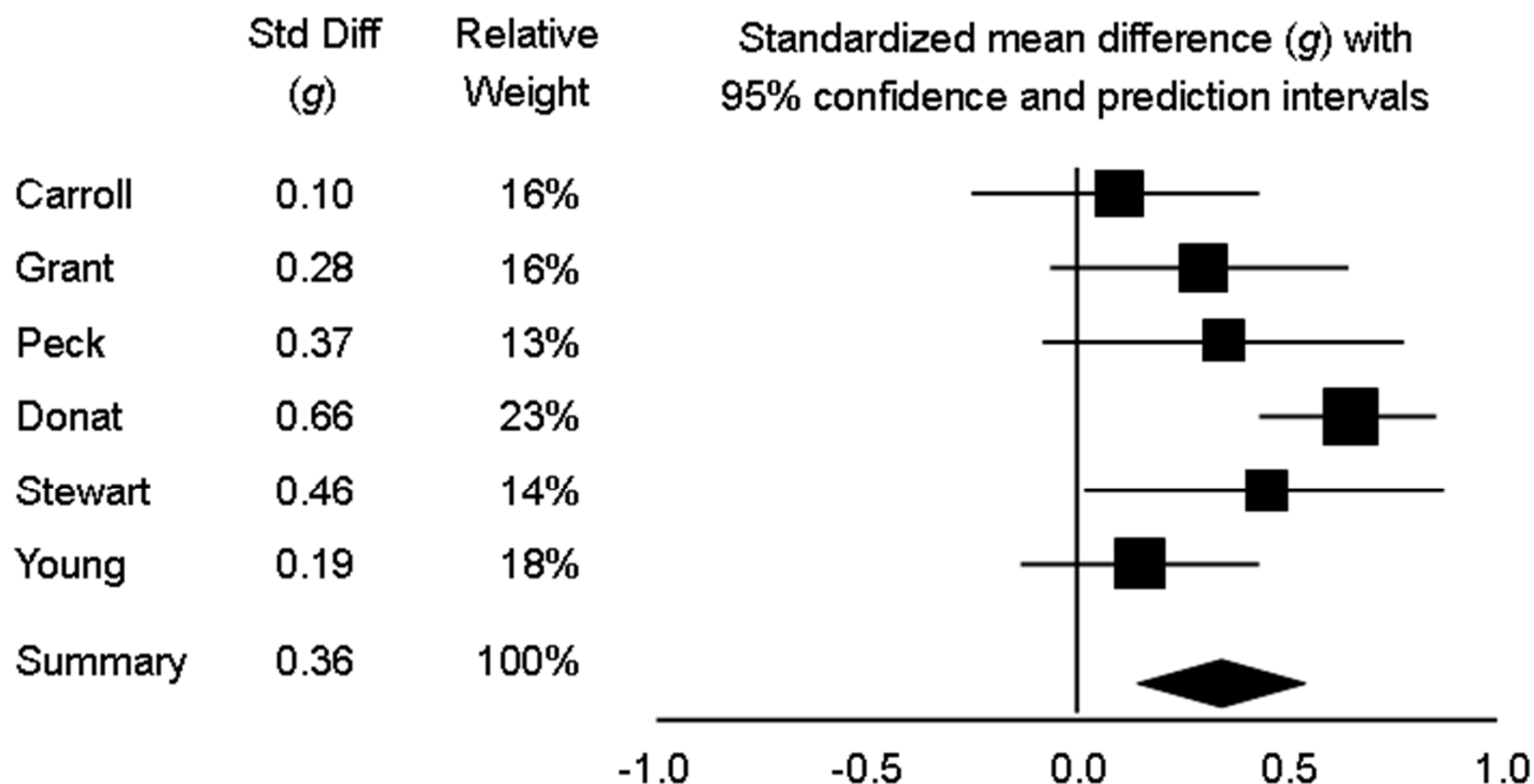
How weights shift

- If within-study variance only, $W=1/V$
- If between-study variance only, $W=1/T^2$
- If both, $W=1/(V+T^2)$

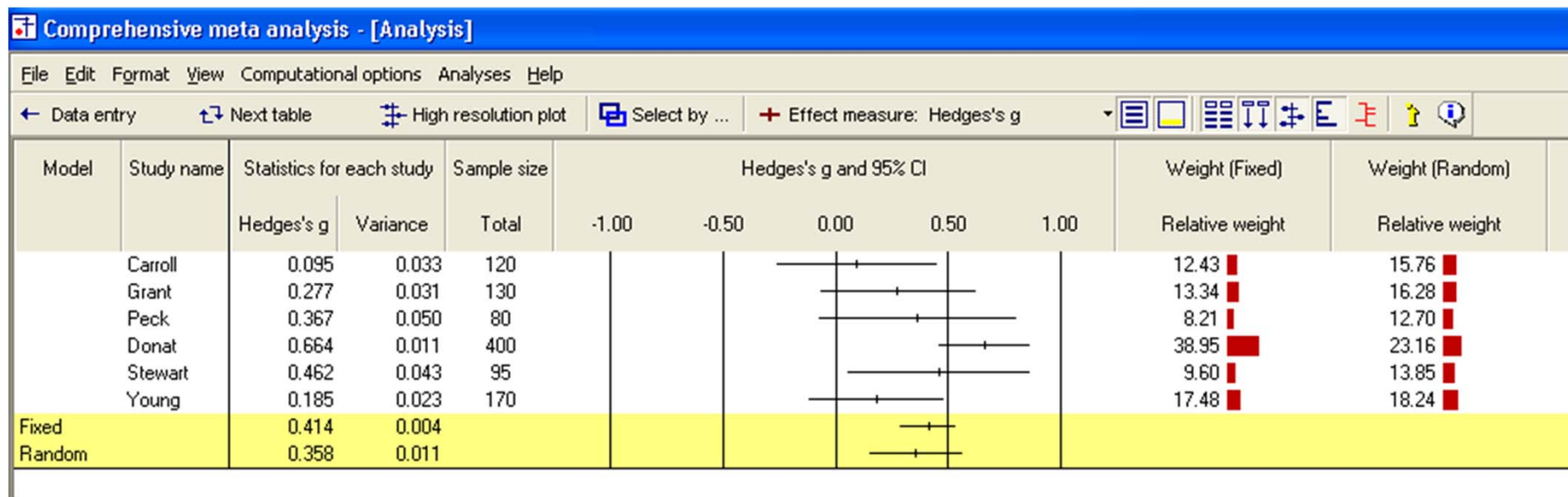
Impact of Intervention (Fixed effect)



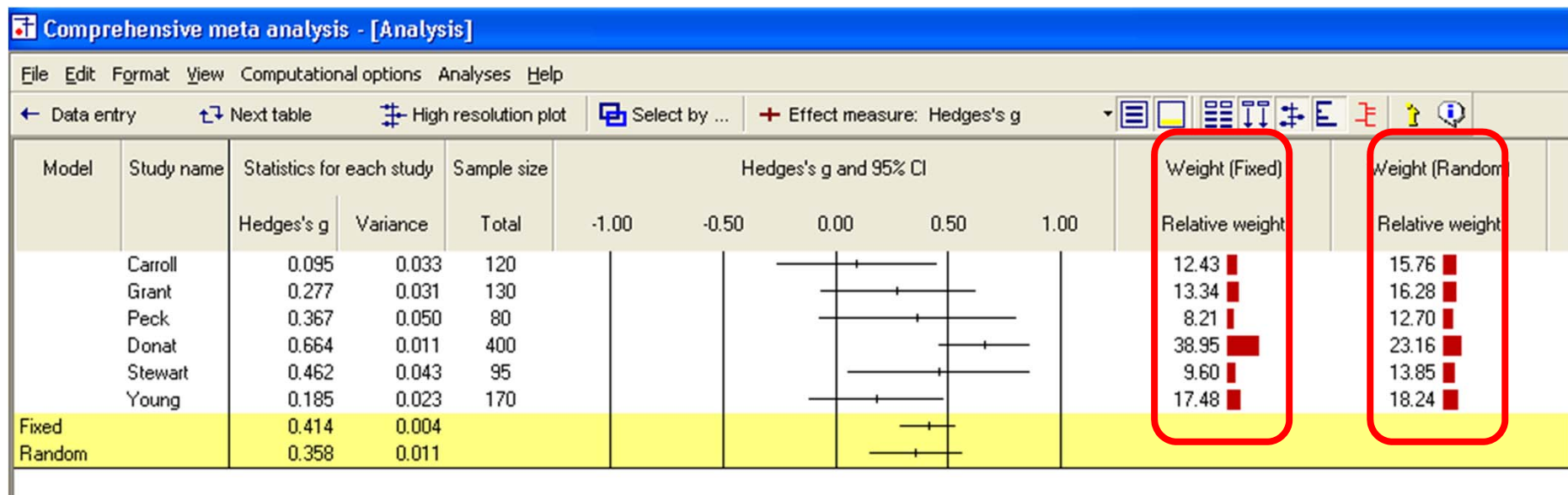
Impact of Intervention (Random effects)



Random vs. Fixed

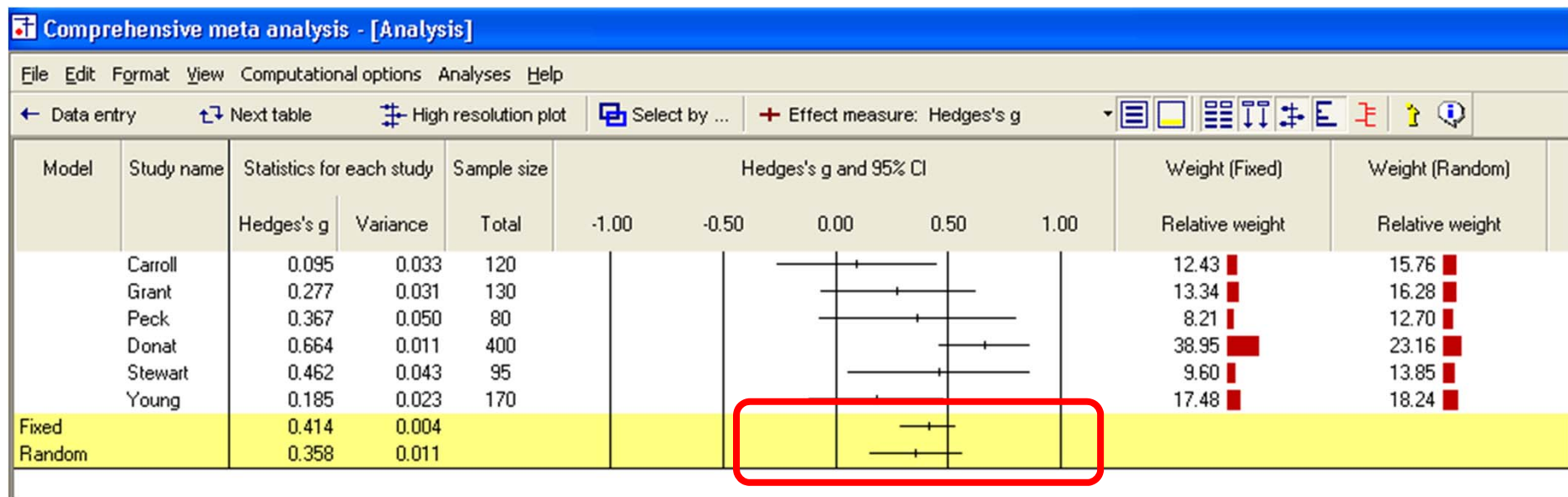


Random vs. Fixed



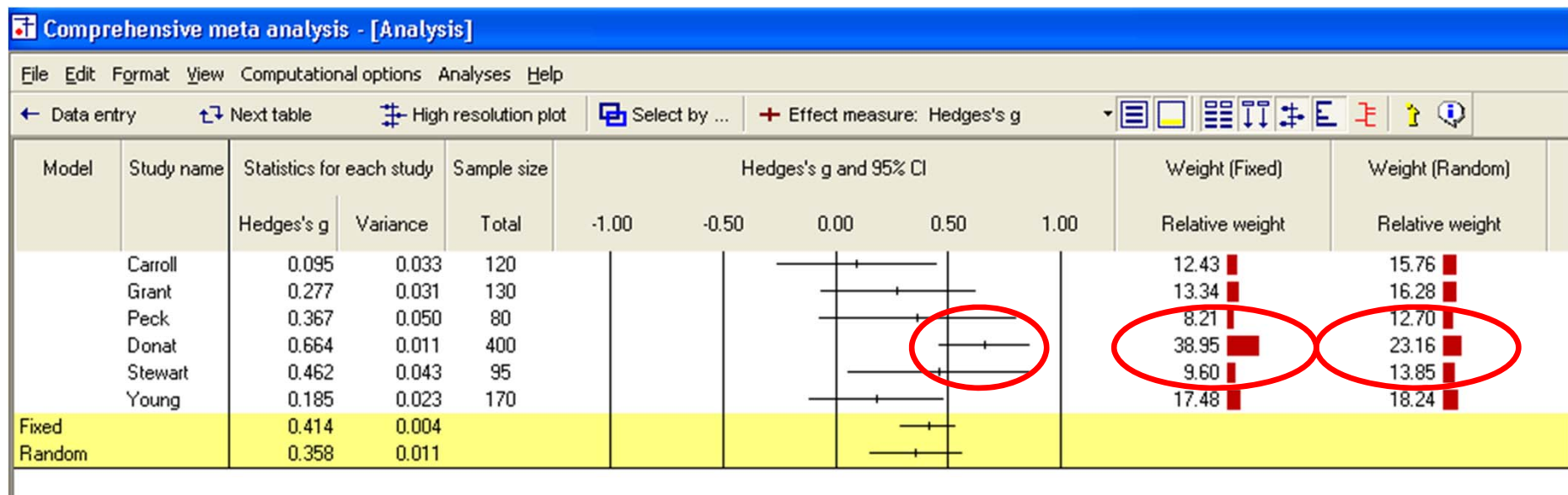
RE weights are more balanced

Random vs. Fixed

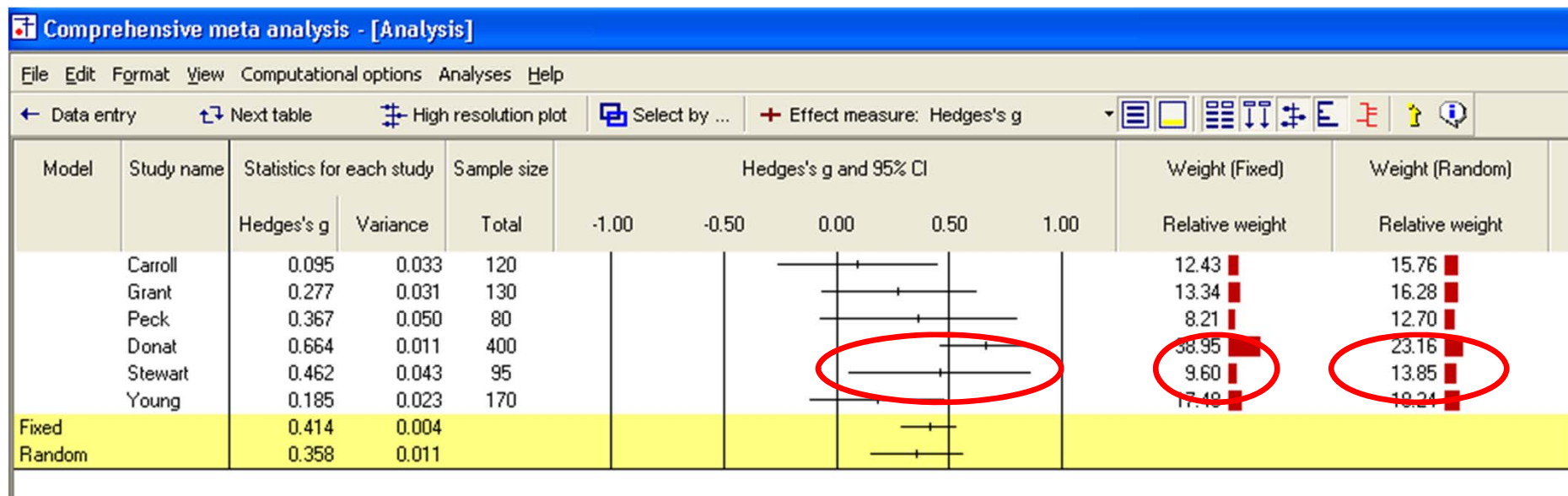


RE confidence interval is wider

Large study has less impact under RE



Small study has more impact under RE



Summary effect

	Hedges's g	Variance	W	T*W
Carroll	0.095	0.033	30.352	2.869
Grant	0.277	0.031	32.568	9.033
Peck	0.367	0.050	20.048	7.349
Donat	0.664	0.011	95.111	63.190
Stewart	0.462	0.043	23.439	10.824
Young	0.185	0.023	42.698	7.906
	0.414	0.004	244.215	101.171

FE

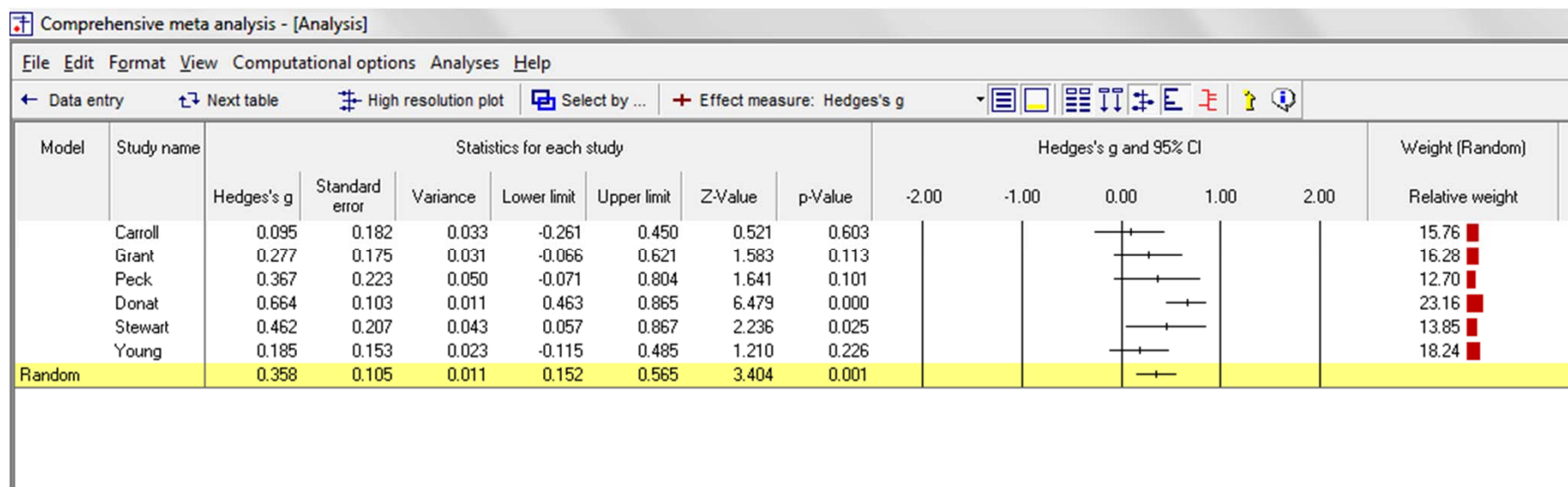
$$\frac{101.171}{244.215} = 0.414$$

Variance of summary effect

	Hedges's g	Variance	W	T*W
Carroll	0.095	0.033	30.352	2.869
Grant	0.277	0.031	32.568	9.033
Peck	0.367	0.050	20.048	7.349
Donat	0.664	0.011	95.111	63.190
Stewart	0.462	0.043	23.439	10.824
Young	0.185	0.023	42.698	7.906
	0.414	0.004	244.215	101.171

FE

$$\frac{1}{244.215} = 0.004$$



RE

Summary effect

Hedges's g	Variance	Tau ² Within	Total Variance	W	T*W
0.095	0.033	0.037	0.070	14.233	1.345
0.277	0.031	0.037	0.068	14.702	4.078
0.367	0.050	0.037	0.087	11.469	4.204
0.664	0.011	0.037	0.048	20.909	13.892
0.462	0.043	0.037	0.080	12.504	5.774
0.185	0.023	0.037	0.061	16.466	3.049
0.358	0.011			90.284	32.342

RE

$$\frac{32.342}{90.284} = 0.358$$

Variance of summary effect

Hedges's g	Variance	Tau ² Within	Total Variance	W	T*W
0.095	0.033	0.037	0.070	14.233	1.345
0.277	0.031	0.037	0.068	14.702	4.078
0.367	0.050	0.037	0.087	11.469	4.204
0.664	0.011	0.037	0.048	20.909	13.892
0.462	0.043	0.037	0.080	12.504	5.774
0.185	0.023	0.037	0.061	16.466	3.049
0.358	0.011			90.284	32.342

RE

$$\frac{1}{90.284} = 0.011$$

Why does it matter?

- One matches the sampling
- One does not
- Wrong model yields incorrect weights
- Estimate of mean is wrong
- Estimate of CI is wrong

Fixed vs. Random

- MUST choose based on sampling model
- The meaning of the ES is different
- Relative weights are closer under RE (effect size will shift)
- Absolute weights are smaller under RE (CI will become wider)
- p -value will change (less significant in long run but can go either way)

Test of null

Fixed

$$Z = \frac{M}{SE_M}$$

One
source
of error



Random

$$Z^* = \frac{M^*}{SE_{M^*}}$$

Two
sources
of error

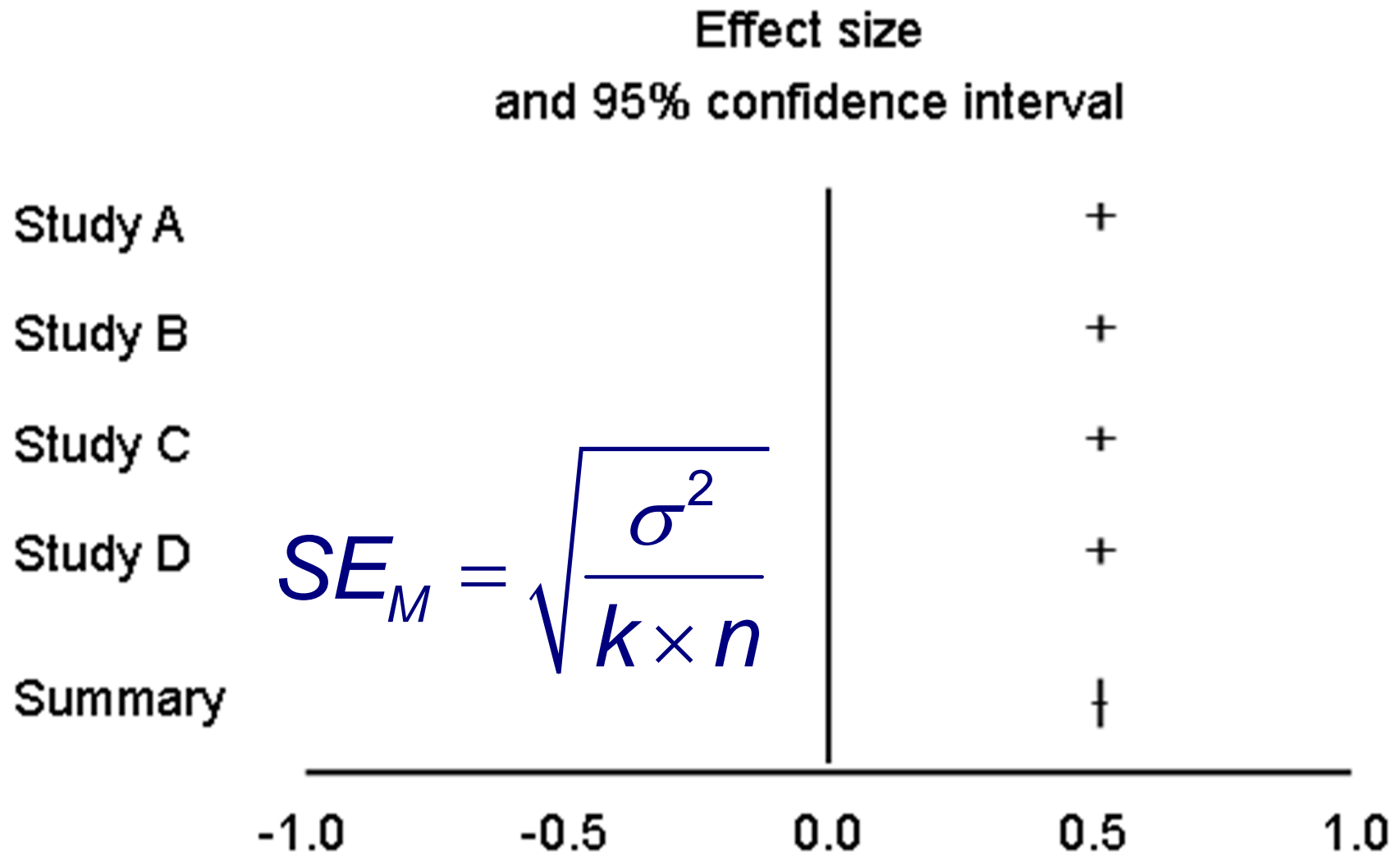


Question!

Suppose we had four studies, each with $N = 1,000,000$, and a true (mean) effect size of 0.50. Under the two models,

- What would the forest plot look like?
- What would the diamond look like?

Fixed-effect model



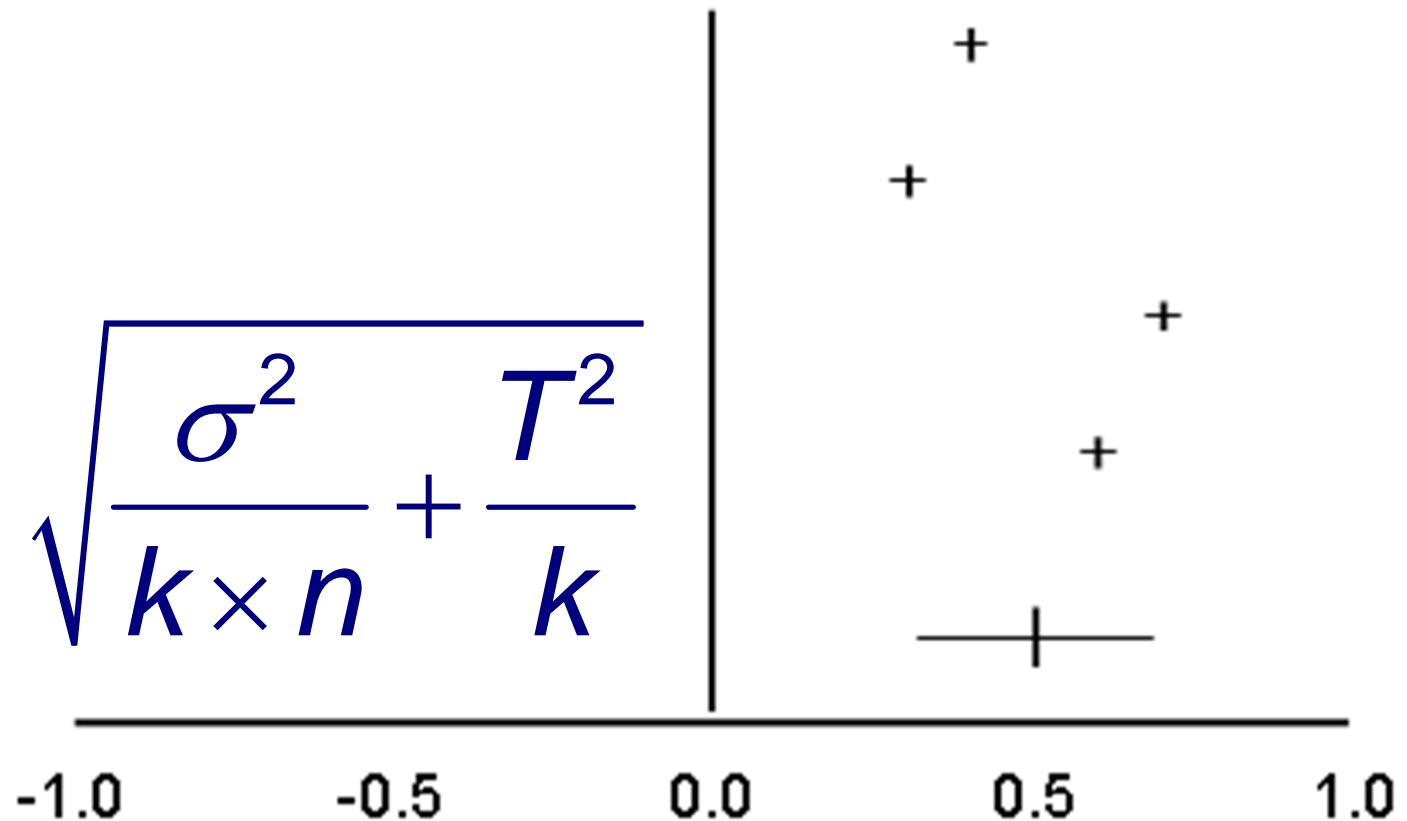
Random-effects model

Effect size
and 95% confidence interval

Study A

Study B

$$SE_{M^*} = \sqrt{\frac{\sigma^2}{k \times n} + \frac{T^2}{k}}$$



Statistical power, Fixed-effect

$$Power = f\left(\frac{d}{SE_d}\right)$$

$$SE_d = \sqrt{\frac{\sigma^2}{k \times n}}$$

Statistical power, Random-effects

$$Power = f\left(\frac{d}{SE_d}\right)$$

$$SE_d = \sqrt{\frac{\sigma^2}{k \times n} + \frac{\tau^2}{k}}$$

Need to know the source of the variance

- If one source, then error is V/n
- If two sources, then error is $V/n + T^2/k$

Which model should we use?

- Base decision on the model that matched the way the data were collected
- *Not* on test of homogeneity



What you may hear

- Fixed-effect is simple model
- Random-effects is more complicated

Actually

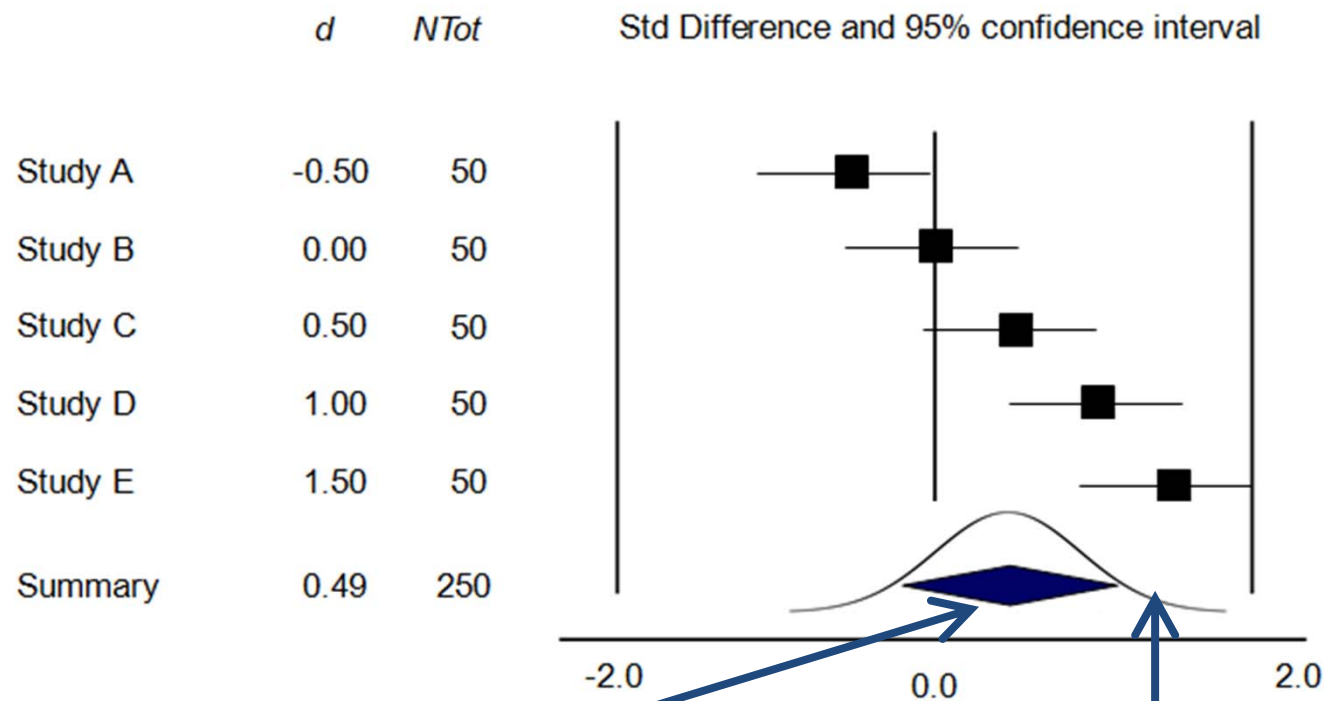
- Fixed-effect is more restricted model
- Random-effects makes less assumptions

An alternate view

- Random-effects model only makes sense if we have a clear picture of the sampling frame
- Otherwise, we should report the mean and CI for the studies in our sample without attempt to generalize to a larger universe
- This is a fixed-effects analysis (in the plural) where “fixed” means “set” rather than “common”

Prediction intervals

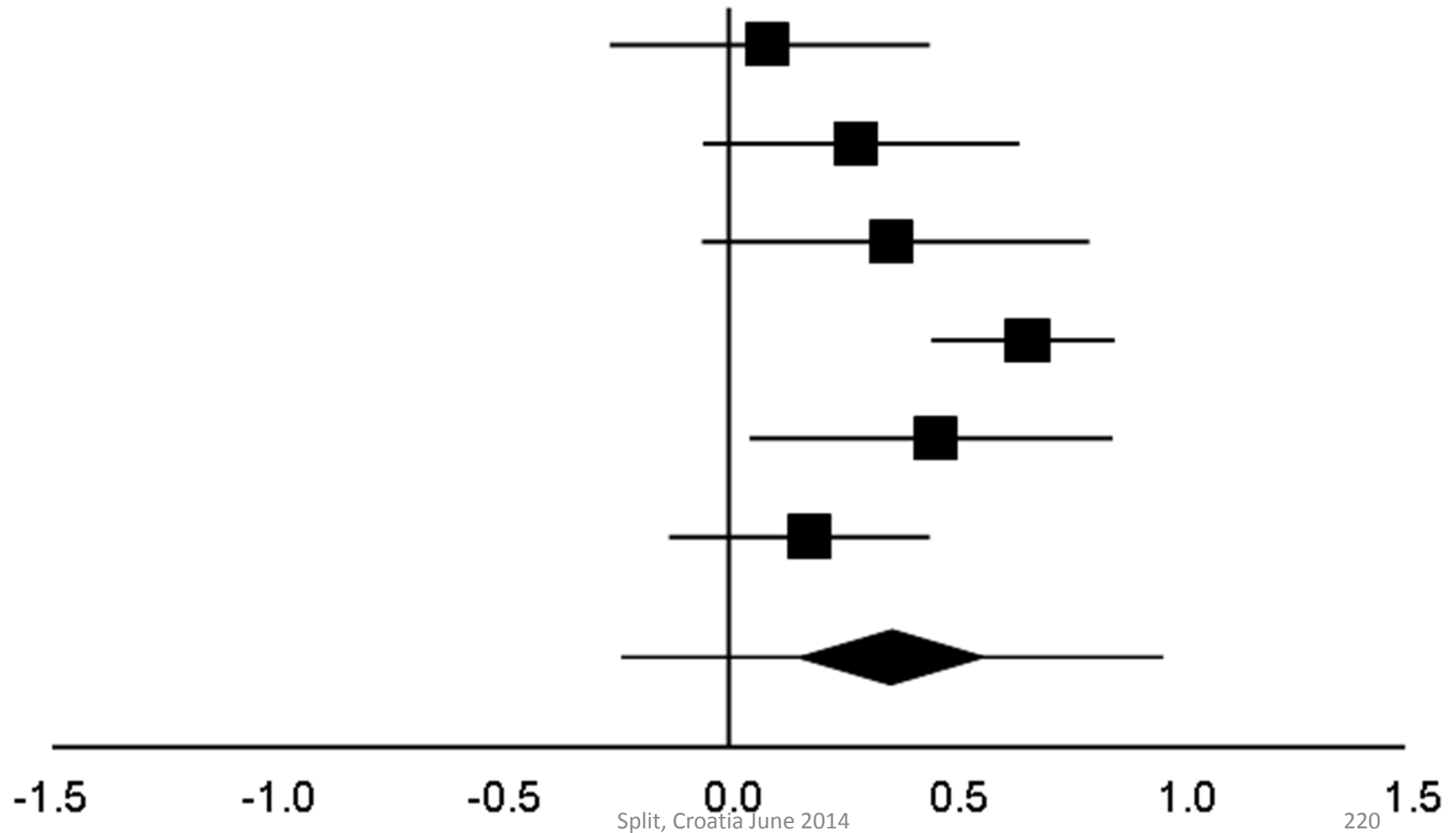
Meta-analysis with heterogeneous effects $k = 5$



Precision of the
mean effect

Dispersion of the
individual effects

Confidence and Prediction



Confidence Interval

- Measure of *precision*
- Range for the true value of the mean
- Analogous to standard error
- Applies to fixed or random effects model

Prediction Interval

- Measure of *dispersion*
- Range for the true effect in different studies
- Analogous to standard deviation

Confidence Interval vs. Prediction Interval

