

Fixed Effects Meta-Analysis and Homogeneity Evaluation

Jeff Valentine
University of Louisville

Campbell Collaboration Annual
Colloquium
Oslo 2009

Inverse-variance Weighted Average

- All effect sizes are not created equal
 - We like effects from big samples much more than effects from small samples
 - Therefore, we weight studies to give preference to larger samples
 - Weight by the inverse of the variance of the effect size
 - For d , inverse variance is

$$1/s^2 = w = \frac{2n_1n_2(n_1 + n_2)}{2(n_1 + n_2)^2 + n_1n_2d^2}$$

- For z_r , inverse variance is $n-3$ ($n = \#$ of pairs of scores)
 - correlations are transformed to z for analysis (see Cooper; Lipsey & Wilson)

Example: Computing Weights

Study	r_i	z_i	Number of pairs of scores	Weight
A	.00	.00	56	53
B	.17	.172	23	20
C	.28	.29	33	30
D	.44	.47	10	7
E	.67	.81	27	24

Example: Computing Weights

Study	d_i	n_1	n_2	$2n_1n_2$	n_1+n_2	Numerator (Col. 5 x Col. 6)	$2(n_1+n_2)^2$	$n_1n_2d^2$	Denominator (Col. 8 + Col. 9)	Weight (Col. 7 ÷ Col. 10)
A	-.33	15	15							
B	.06	60	60							
C	.29	20	20							
D	.45	10	10							
E	-.04	40	40							

Example: Computing Weights

Study	d_i	n_1	n_2	$2n_1n_2$	n_1+n_2	Numerator (Col. 5 x Col. 6)	$2(n_1+n_2)^2$	$n_1n_2d^2$	Denominator (Col. 8 + Col. 9)	Weight (Col. 7 ÷ Col. 10)
A	-.33	15	15	450						
B	.06	60	60	7200						
C	.29	20	20	800						
D	.45	10	10	200						
E	-.04	40	40	3200						

Example: Computing Weights

Study	d_i	n_1	n_2	$2n_1n_2$	n_1+n_2	Numerator (Col. 5 x Col. 6)	$2(n_1+n_2)^2$	$n_1n_2d^2$	Denominator (Col. 8 + Col. 9)	Weight (Col. 7 ÷ Col. 10)
A	-.33	15	15	450	30					
B	.06	60	60	7200	120					
C	.29	20	20	800	40					
D	.45	10	10	200	20					
E	-.04	40	40	3200	80					

Example: Computing Weights

Study	d_i	n_1	n_2	$2n_1n_2$	n_1+n_2	Numerator <small>(Col. 5 x Col. 6)</small>	$2(n_1+n_2)^2$	$n_1n_2d^2$	Denominator <small>(Col. 8 + Col. 9)</small>	Weight <small>(Col. 7 ÷ Col. 10)</small>
A	-.33	15	15	450	30	13500				
B	.06	60	60	7200	120	864000				
C	.29	20	20	800	40	32000				
D	.45	10	10	200	20	4000				
E	-.04	40	40	3200	80	256000				

Example: Computing Weights

Study	d_i	n_1	n_2	$2n_1n_2$	n_1+n_2	Numerator <small>(Col. 5 x Col. 6)</small>	$2(n_1+n_2)^2$	$n_1n_2d^2$	Denominator <small>(Col. 8 + Col. 9)</small>	Weight <small>(Col. 7 ÷ Col. 10)</small>
A	-.33	15	15	450	30	13500	1800			
B	.06	60	60	7200	120	864000	28800			
C	.29	20	20	800	40	32000	3200			
D	.45	10	10	200	20	4000	800			
E	-.04	40	40	3200	80	256000	12800			

Example: Computing Weights

Study	d_i	n_1	n_2	$2n_1n_2$	n_1+n_2	Numerator <small>(Col. 5 x Col. 6)</small>	$2(n_1+n_2)^2$	$n_1n_2d^2$	Denominator <small>(Col. 8 + Col. 9)</small>	Weight <small>(Col. 7 ÷ Col. 10)</small>
A	-.33	15	15	450	30	13500	1800	25.50		
B	.06	60	60	7200	120	864000	28800	12.96		
C	.29	20	20	800	40	32000	3200	33.64		
D	.45	10	10	200	20	4000	800	20.25		
E	-.04	40	40	3200	80	256000	12800	2.56		

Example: Computing Weights

Study	d_i	n_1	n_2	$2n_1n_2$	n_1+n_2	Numerator <small>(Col. 5 x Col. 6)</small>	$2(n_1+n_2)^2$	$n_1n_2d^2$	Denominator <small>(Col. 8 + Col. 9)</small>	Weight <small>(Col. 7 ÷ Col. 10)</small>
A	-.33	15	15	450	30	13500	1800	25.50	1824.5	
B	.06	60	60	7200	120	864000	28800	12.96	28812.96	
C	.29	20	20	800	40	32000	3200	33.64	3233.64	
D	.45	10	10	200	20	4000	800	20.25	820.25	
E	-.04	40	40	3200	80	256000	12800	2.56	12802.56	

Example: Computing Weights

Study	d_i	n_1	n_2	$2n_1n_2$	n_1+n_2	Numerator (Col. 5 x Col. 6)	$2(n_1+n_2)^2$	$n_1n_2d^2$	Denominator (Col. 8 + Col. 9)	Weight (Col. 7 ÷ Col. 10)
A	-.33	15	15	450	30	13500	1800	25.50	1824.5	7.40
B	.06	60	60	7200	120	864000	28800	12.96	28812.96	29.99
C	.29	20	20	800	40	32000	3200	33.64	3233.64	9.90
D	.45	10	10	200	20	4000	800	20.25	820.25	4.88
E	-.04	40	40	3200	80	256000	12800	2.56	12802.56	20.00

Computing a Weighted Average Effect Size

- Conceptually no different from any other weighted average

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

Here,

$$\overline{d} = .05$$

Study	ES_i	w_i	$w_i^*ES_i$
A	-.33	7.40	-2.44
B	.06	29.99	1.80
C	.29	9.90	2.87
D	.45	4.88	2.19
E	-.04	20.0	-.80
	$\Sigma =$	72.15	3.62

Computing a 95% Confidence Interval for Average Effect Size

- Just like every other statistic you've looked at, the confidence interval (CI) is based on a *standard error* and *critical test value*

- The standard error of the mean effect size is given by the formula $SE_{\bar{ES}} = \sqrt{\frac{1}{\sum w_i}}$

- We use the distribution of z to test the significance of effect sizes in meta-analysis

- At the .05 significance level (i.e., 95% confidence level), two-tailed, $z = 1.96$

- From our example, $\sum w_i = 72.15$, therefore

$$SE_d = \sqrt{\frac{1}{72.15}} = .12$$

$$95\% CI = 1.96(.12) = .23$$

$$\bar{d} = .05 \pm .23$$

13

Relationship between CIs and Statistical Testing

- If the CI for the mean ES does not include zero, can reject the null hypothesis of no population effect
- Here, the CI ranged from $-.18$ to $+.28$, so cannot reject H_0 of no population effect
- Also can do a direct test

$$z = \frac{\bar{ES}}{SE_{\bar{ES}}} = \frac{.05}{.12} = .4167, p = .68$$

- Exact p-values can be obtained from statistical tables, online calculators, and MS Excel (and presumably other spreadsheets)

One more thing: Relative Weights

- Relative weights are a good way to see how much influence each study exerts on the average, relative to other studies
 - Relative weight is just the percentage of the sum of the weights that can be attributed to any individual study

Study	d_j	Total n	Raw weight	Relative weight
A	-.33	30	7.40	10.3%
B	.06	120	29.99	45.6%
C	.29	40	9.90	13.7%
D	.45	20	4.88	6.7%
E	-.04	80	20.00	27.7%

Testing Effect Sizes for Homogeneity

- Homogeneity
 - “composed of parts or elements that are all of the same kind; not heterogeneous” – dictionary.com
 - Statistically, asks the question “Do these studies appear to be estimating the same underlying population parameter?”
 - Subject-level sampling error is the only reason that the studies arrived at different effect sizes

More on Heterogeneity

- Studies are based on samples of participants
 - Sample statistics vary due to *sampling error*
 - True even if all samples are estimating the same population quantity



A Scenic Route

- Recall that to compute a variance of a set of scores

$$s^2 = \frac{\Sigma(Y - \bar{Y})^2}{n - 1}$$

- And that to compute a between-groups sum of squares in ANOVA

$$SS_{Bet} = \Sigma n_j (\bar{Y}_j - \bar{Y}_T)^2$$

where

n_j = # in group j

\bar{Y}_j = mean of group j

\bar{Y}_T = grand mean

Statistical Test of Homogeneity

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

where

ES_i = each individual effect size

\overline{ES} = the weighted mean effect size

w_i = the individual weight for ES_i

Statistical Test of Homogeneity

Null hypothesis is that $\delta_1 = \delta_2 = \delta_3 = \dots = \delta_i$

That is, that all sampled effect sizes are estimating the same population parameter δ (delta)

Another way of thinking about this:

We don't expect effect sizes to have the exactly the same values, due to sampling error. Do the effect sizes vary more than we would expect given sampling error?

Note that this is just a weighted sums of squares

Computing the Homogeneity Test

Study	ES_i	ES_{avg}	w_i	$ES-ES_{avg}$	$(ES-ES_{avg})^2$	$w_i (ES-ES_{avg})^2$
A	-.33	.05	7.40	-.38		
B	.06	.05	29.99	.01		
C	.29	.05	9.90	.24		
D	.45	.05	4.88	.4		
E	-.04	.05	20.00	-.09		
						$\Sigma =$

Computing the Homogeneity Test

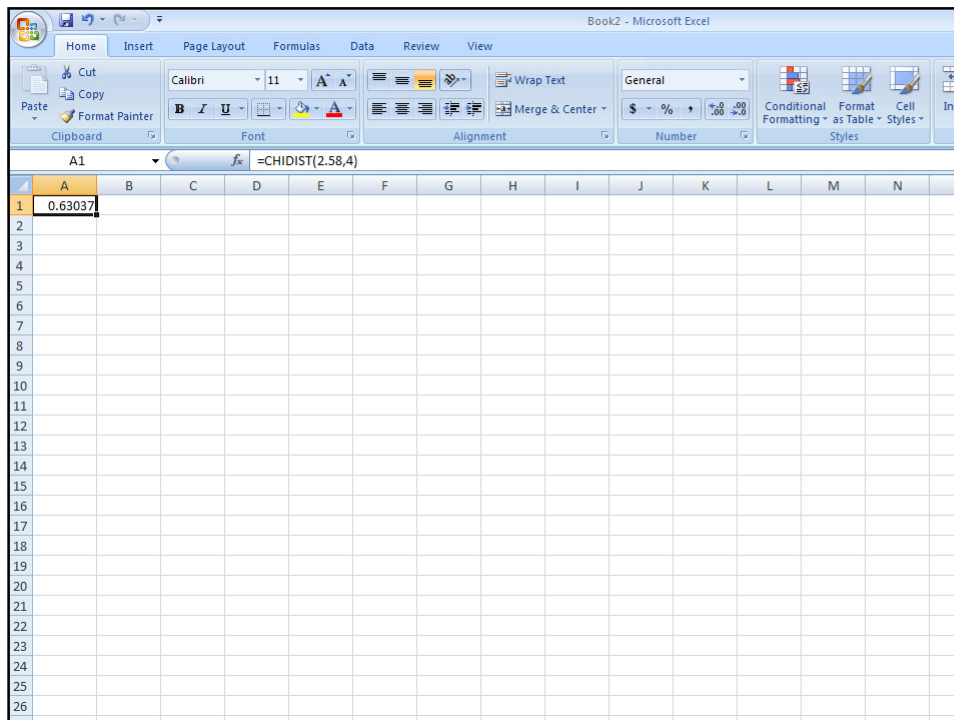
Study	ES_i	ES_{avg}	w_i	$ES-ES_{avg}$	$(ES-ES_{avg})^2$	$w_i (ES-ES_{avg})^2$
A	-.33	.05	7.40	-.38	.144	
B	.06	.05	29.99	.01	.0001	
C	.29	.05	9.90	.24	.058	
D	.45	.05	4.88	.4	.16	
E	-.04	.05	20.00	-.09	.008	
						$\Sigma =$

Computing the Homogeneity Test

Study	ES_i	ES_{avg}	w_i	$ES-ES_{avg}$	$(ES-ES_{avg})^2$	$w_i (ES-ES_{avg})^2$
A	-.33	.05	7.40	-.38	.144	1.07
B	.06	.05	29.99	.01	.0001	.003
C	.29	.05	9.90	.24	.058	.57
D	.45	.05	4.88	.4	.16	.78
E	-.04	.05	20.00	-.09	.008	.16
						$\Sigma = 2.58$

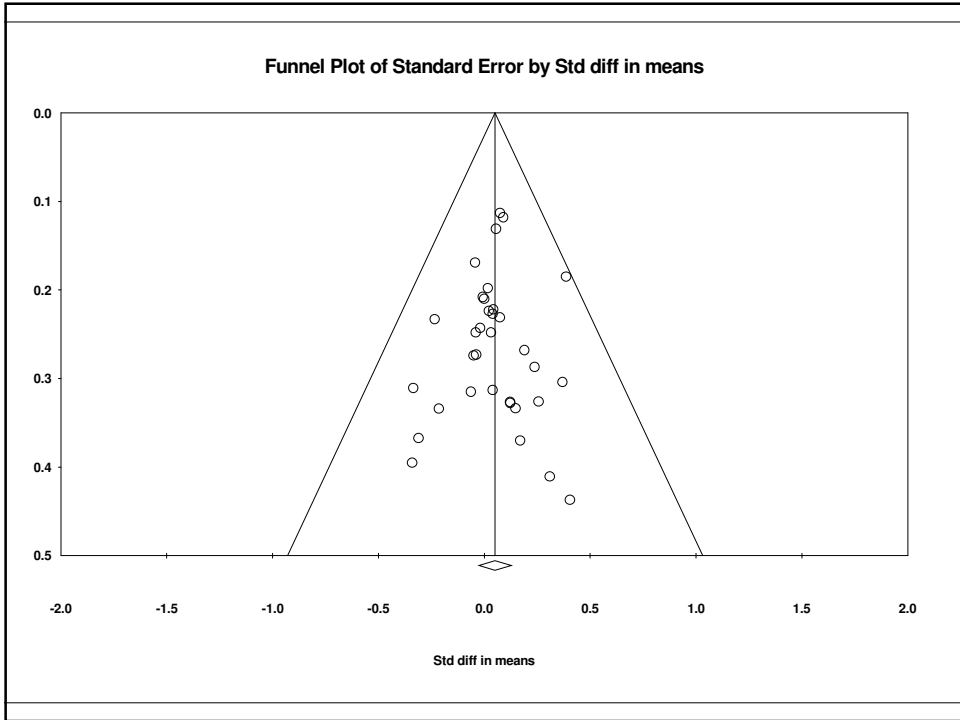
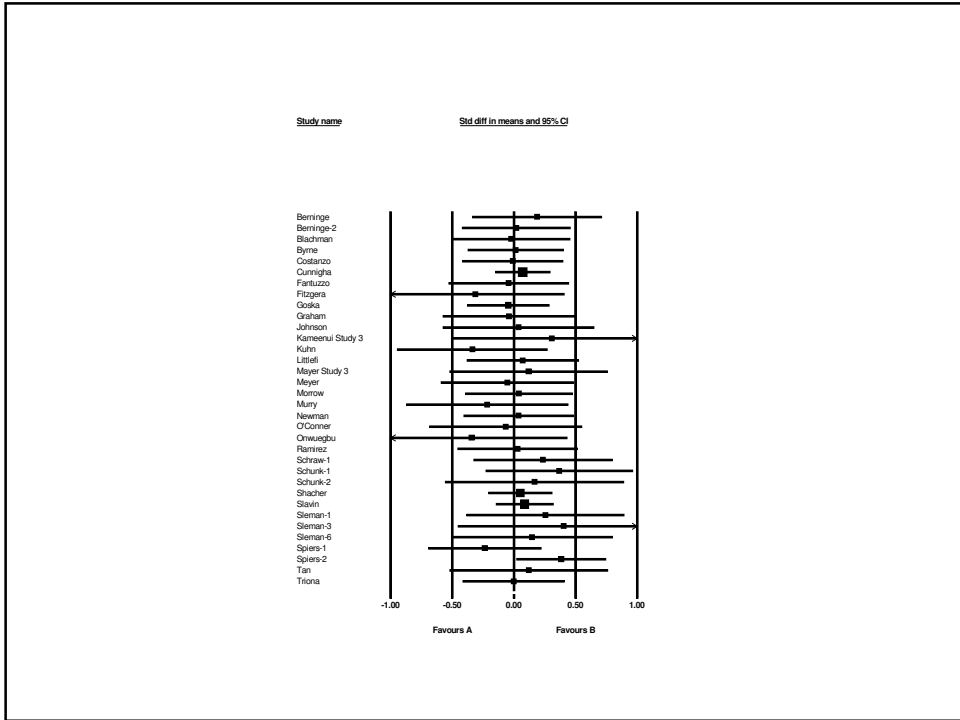
How to Evaluate Q

- Microsoft Excel
 - Use the function =chidist(value, df) to generate an exact p-value for a specific chi square value at given df
 - =chidist(2.58,4) yields p = .63
- Statistical textbooks
 - Often have statistical tables in the Appendix that can be used to generate more or less precise p-values
- Online calculators
 - Are also available
- Specialized software for meta-analysis will also compute exact p-values
 - Comprehensive Meta-Analysis
 - TrialStat (?)
 - RevMan (Cochrane)



Evaluating Q for the Example

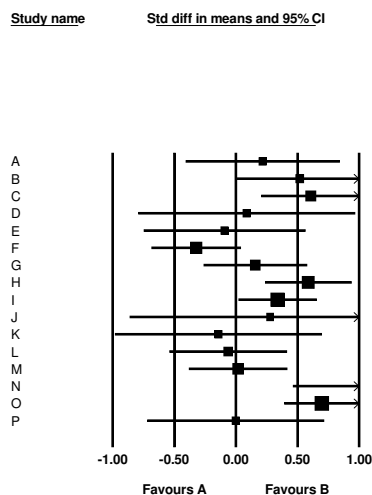
- At 4 df and $\alpha = .05$, critical value of $\chi^2 = 9.49$, so we cannot reject H_0
 - Exact p -value is .63
- “The test of homogeneity was not statistically significant, $Q(4) = 2.58$, $p = .63$, suggesting that the studies are all estimating the same population parameter”
- If statistically significant, we would say that the data are heterogeneous

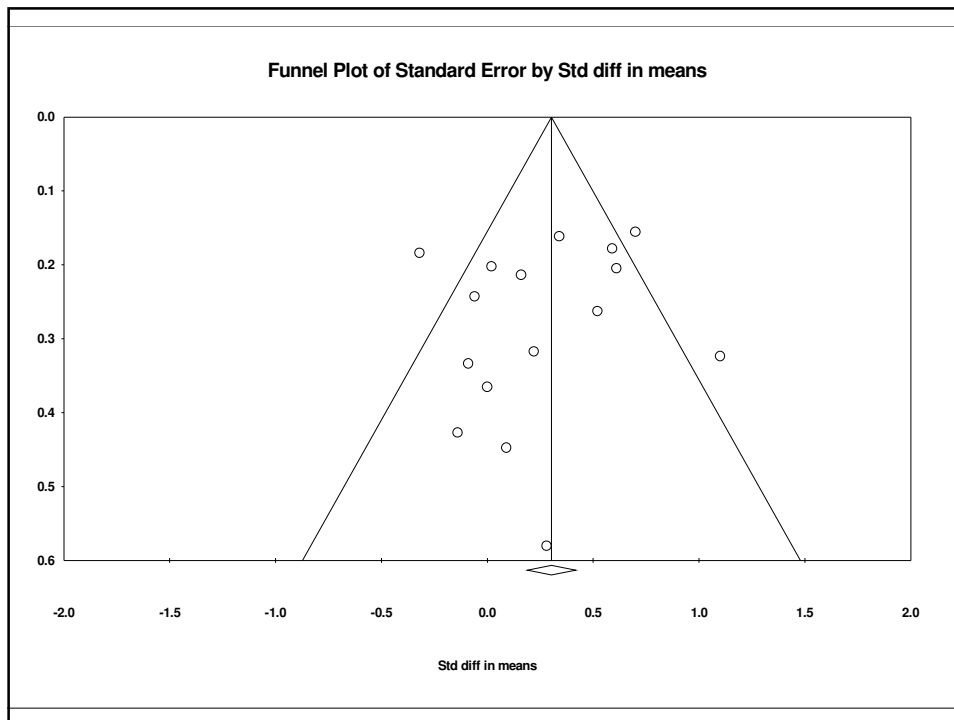


What would a Heterogeneous Distribution of Effect Sizes Look Like?

- For the next example,
 - $Q(15) = 37.80, p = .001$

“The test of homogeneity was statistically significant, $Q(15) = 37.80, p = .001$, suggesting the studies are not estimating the same population parameter.”





Another Way to Think About Heterogeneity

- The homogeneity test can have low statistical power
 - Or even “too much” power (if there is such a thing!)
- Often see an “effect size” describing the magnitude of heterogeneity (i.e., “how much” heterogeneity)
 - Referred to as I^2

$$I^2$$

Computed as

$$I^2 = \frac{Q - (k - 1)}{Q} \times 100$$

- By convention, negative values of I^2 are set to 0
- Higgins et al. (2003) suggest
 - $I^2 = 25\%$ is a small degree of heterogeneity
 - $I^2 = 50\%$ is a moderate degree of heterogeneity
 - $I^2 = 75\%$ is a large degree of heterogeneity

Note that the homogeneity test can be non-significant with a large I^2 (suggesting low power), significant with a small I^2 (suggesting "too much" power)

$Q(4) = 2.40, p = .63, I^2 = 0\%$

$Q(15) = 37.80, p = .001, I^2 = 60\%$

Implication of the Homogeneity Test

- Not statistically significant
 - Implies that the individual ESs vary only due to random sampling error in the individual studies
 - They have different estimates of the population parameter only because they drew a different sample of participants
 - Often used to argue that one shouldn't explore potential sources of variability
 - Analogous to the recommendation for not conducting post-hoc tests in a multi-group ANOVA when F is not statistically significant
- Statistically significant
 - Suggests that factors in addition to sampling error may be contributing to differences in observed effect sizes
 - Often used as a justification to explore sources of variability in effect sizes
 - Sometimes used to justify the use of a *random effects* model

Fixed Effects Meta-Analysis

- Discussion so far has been about *fixed effects* meta-analysis
 - Assumption is that all studies are estimating the same population parameter
 - Inferences should only be extended to situations that are highly similar to those observed in the studies themselves
 - Less formally – a fixed effects meta-analysis can tell you what *these particular studies* say about the relationship in question
 - However, we are virtually always interested in generalizing beyond the particulars of the studies in any given meta-analysis
 - The only thing affecting uncertainty about the population mean is sample size (as the number of total participants across studies goes up, uncertainty goes down)
 - Remember, the SE of an effect size is

$$SE_{\bar{ES}} = \sqrt{\frac{1}{\sum w_i}}$$

Choosing Between Fixed and Random Effects Models

- The choice between fixed and random effects models has important implications
 - RE models are addressed in detail in another training session
 - Most important implication
 - RE models generally have lower statistical power than FE models

Characteristics Suggesting a Fixed Effects Model Might be Appropriate

- If studies are *highly* similar to one another
 - Nature of the intervention, sample, etc.
- If test of homogeneity is not statistically significant
 - Be aware of low statistical power though

NB – for most C2 topics, studies will not be highly similar to one another

NB #2 – The RE model reduces to the FE model if the assumptions of the FE model are exactly met

Comparison of Results: Fixed vs. Random Effects Analyses

	Average Effect Size	Lower Limit of CI	Upper Limit of CI
Fixed	.30	.18	.42
Random	.27	.08	.46

Recap: Implications of the Choice of Analysis Models

	Inferences	Effect Size	Confidence Interval	Statistical Power
Fixed	Conditional on observed studies – generalize only to studies highly like those in the analysis	Weighted average of observed effects; weights computed assuming subject-level sampling error explains all of the study-to-study variability	"Sample size" (through effect size precision estimates) are the only influence on width of CI's – regardless, the larger the total number of observations the smaller the CI	Usually higher than random effects
Random	Unconditional on observed studies – generalize to a broader population of studies	Weighted average of observed effects; weights computed taking into account subject-level sampling error <i>and</i> study-to-study variability	"Sample size" and study-to-study variability influence width of CI's – random effects CI's will never be smaller than their fixed effects counterparts	Usually lower than fixed effects

Choosing Between Fixed and Random Effects Models

- Ultimately, there is no "right" answer
 - Parallels a debate that has been going on in statistics for 70 years
- Hedges & Vevea (1998)
 - Fixed effects model – allows inferences to the observed studies only
 - Random effects model – allows inferences to a population of studies from which the observed studies were randomly sampled
 - So base the decision on where you would like to extend inferences
- The empirical approach
 - Adopt a fixed effects model if the homogeneity test is not statistically significant
 - Need to be aware of the power of this test though – not always high
- The random approach
 - If the assumptions of the fixed effects model are met, there is *no penalty* for employing the random effects model
 - So employ the random effects model
- The compromise approach
 - Report both fixed and random effects models, let the reader decide which to use

Fixed Effects Meta-Analysis and Homogeneity Evaluation

Jeff Valentine
University of Louisville

Campbell Collaboration Annual
Colloquium
Oslo 2009