

Systematic Reviews

by

Larry V. Hedges

The University of Chicago

WHAT IS A SYSTEMATIC REVIEW?

Systematic reviews summarize evidence in ways that limit bias in:

Assembly of studies (data collection)

Critical appraisal of studies (data evaluation)

Synthesis of evidence across studies (meta-analysis)

As in all scientific work, rigorous methods and clear reporting (transparency) are crucial

WHAT IS A SYSTEMATIC REVIEW?

Feature	Narrative Review	Systematic Review
Question	Broad	Focused
Sources & Search	Not usually specified, possibly biased	Explicit search strategy
Selection	Not usually specified, possibly biased	Criterion-based
Appraisal	Variable	Explicit criteria
Synthesis	Often qualitative	Quantitative
Inferences	Sometimes evidence based	Usually evidence based

FORMULATING QUESTIONS

Well formulated questions should define:

- The patients of interest
- The main interventions under investigation
- The comparison group (intervention)
- The outcomes of interest

STUDY SELECTION

Planning study selection

- Start with a well formulated question
- Selection criteria that fit the clinical question
- Specify types of study designs included
- Specify type and form of publication
- Write a detailed protocol !

STUDY SELECTION

Selecting studies

- Follow the protocol
- Have 2 investigators review each study
- Consider “blinding” study results

STUDY APPRAISAL

- Examine important clinical features
- Evaluate the quality of study methods
- Construct and pretest appraisal forms
- Write a detailed protocol !

SYNTHESIS IN SYSTEMATIC REVIEWS

Synthesis in systematic reviews is a statistical process (meta-analysis)

The results of each study are represented by an index of effect size

These effect size estimates are combined across studies

EFFECT SIZES

Effect sizes represent the *results* of studies in a way that:

- Is comparable across studies
- Does not depend on study design (sample size)
- Is substantively interpretable

EFFECT SIZES

The choice of effect size depends on the

- *measures* of the outcome variable
- *designs* of studies being reviewed
- *statistical analyses* that have been reported

EFFECT SIZES

Most systematic reviews will probably use effect sizes from one of two families

- the *standardized mean difference* family (including the *d*-index)
- the *odds ratio* family (including the risk difference and risk ratio)

STATISTICAL CONSIDERATION

A crucial conceptual distinction is between effect size

- estimates computed from studies
(*sample* effect sizes)
- parameters
(*population* or *true* effect sizes)

We want to infer about effect size parameters using effect size estimates

STANDARDIZED MEAN DIFFERENCE

The standardized mean difference may be appropriate when

- studies use different (continuous) outcome measures (e.g., many PROs)
- study designs compare the mean outcomes in treatment and control groups
- analyses use General Linear Model or t-tests

STANDARDIZED MEAN DIFFERENCE

Population

Sample

Study Data

Means

$$\mu^T$$

$$\mu^C$$

$$\bar{Y}^T \quad \bar{Y}^C$$

SD

$$\sigma$$

$$S$$

Effect Sizes

$$\delta = \frac{\mu^T - \mu^C}{\sigma}$$

$$d = \frac{\bar{Y}^T - \bar{Y}^C}{S}$$

THE ODDS RATIO FAMILY

The odds ratio family of effect sizes may be appropriate when

- studies use a dichotomous outcome measure
- study designs compare the mean outcomes in treatment and control groups
- analyses use chi-square (or *Generalized Linear Model*) tests

THE ODDS RATIO FAMILY

The mean outcome is measured as the proportion of cases having *one* of the two outcomes (the target outcome)

Study data are proportions of the T and C group having the target outcome

Population

π^T π^C

Sample

p^T p^C

THE ODDS RATIO FAMILY

There are several ways to make an effect size by comparing π^T with π^C

Population

Sample

Risk Difference

$$\Delta = \pi^T - \pi^C$$

$$RD = p^T - p^C$$

Risk Ratio

$$\rho = \pi^T / \pi^C$$

$$RR = p^T / p^C$$

Odds Ratio

$$\omega = \frac{\pi^T (1 - \pi^C)}{\pi^C (1 - \pi^T)}$$

$$OR = \frac{p^T (1 - p^C)}{p^C (1 - p^T)}$$

STATISTICAL CONSIDERATIONS

Standard errors express the sampling uncertainty of an effect size estimate

Standard errors depend on a study's sample size

If sample sizes vary across studies, so do standard errors

The standard error can be calculated from a single study

STATISTICAL CONSIDERATIONS

Therefore a confidence interval for the effect size parameter θ can be computed from a single sample effect size estimate T and its standard error

The 95% confidence interval for θ is

$$T - 1.96SE(T) \leq \theta \leq T + 1.96SE(T)$$

STATISTICAL CONSIDERATIONS

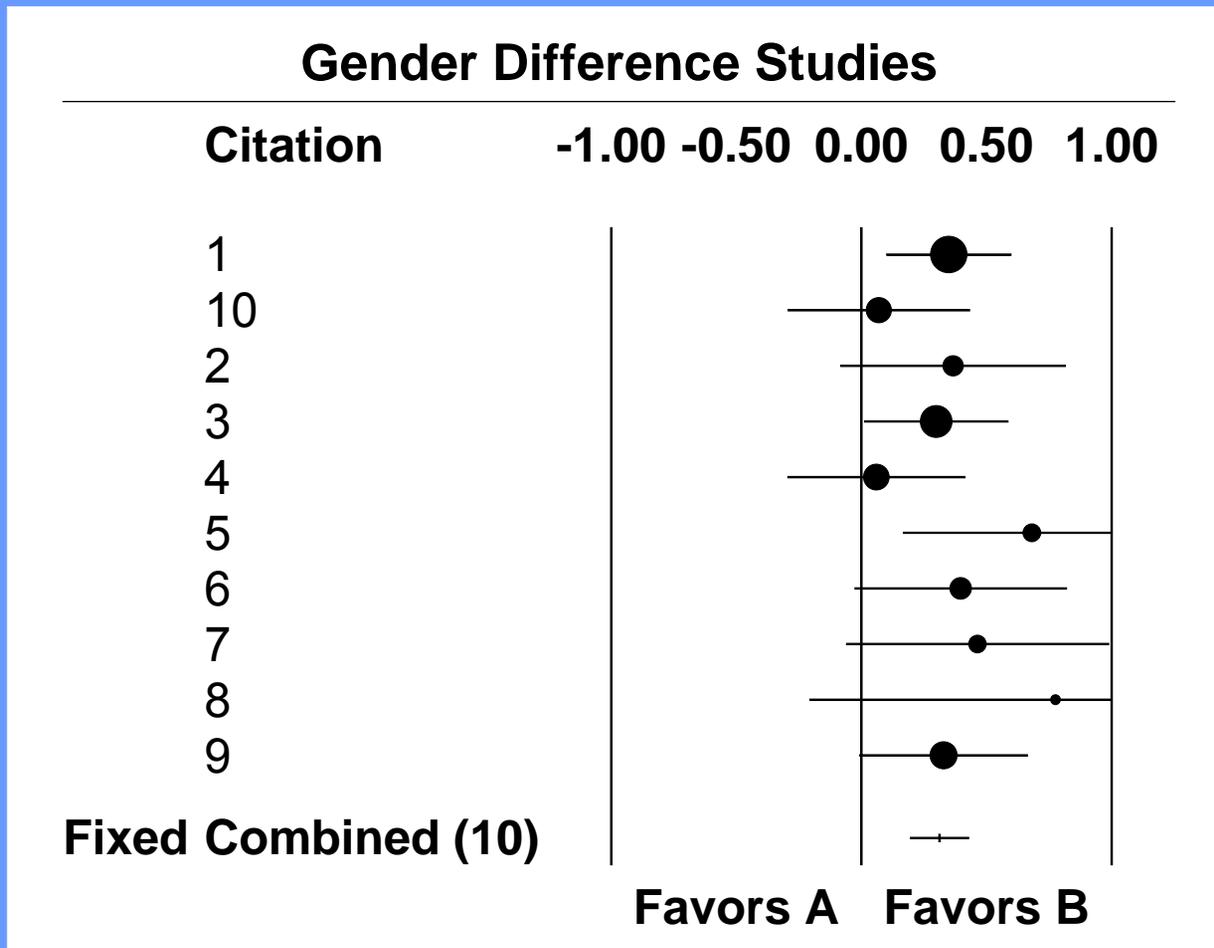
A graph of effect size estimates and their confidence intervals from several studies is called a Forest plot

It displays both effects and their uncertainties

The size of the dot in the center is proportional to the sample size or (inversely) to standard error

STATISTICAL CONSIDERATIONS

Example of a Forest Plot



COMBINING ESTIMATES ACROSS STUDIES

Now suppose we have computed effect size estimates from k studies, call them

$$T_1, T_2, \dots, T_k$$

Call their variances (squares of their SE's)

$$v_1, v_2, \dots, v_k$$

Call the population effect sizes

$$\theta_1, \theta_2, \dots, \theta_k$$

COMBINING ESTIMATES ACROSS STUDIES

To combine these estimates across studies it makes sense to average them

But

All studies do not provide estimates with the same precision

Therefore it makes sense to give more weight to the more precise studies

COMBINING ESTIMATES ACROSS STUDIES

There are two different inference models for combining, with slightly different objectives

Conditional models try to estimate the mean effect size of the studies that are observed

Unconditional models try to estimate the mean effect size of the population of studies from which the observed studies are a sample

COMBINING ESTIMATES ACROSS STUDIES

These two inference models lead to slightly different statistical procedures for combining effect size estimates across studies

Conditional models lead to *fixed effects* statistical procedures

Unconditional models lead to *random effects* statistical procedures

FIXED EFFECTS PROCEDURES

In *fixed effects procedures*,

the goal is to estimate the mean θ_0 of the effect size parameters in the studies that are observed

Between-study variation in effect sizes has no impact on the weights or the uncertainty of this mean

FIXED EFFECTS PROCEDURES

This suggests a weighted mean like

$$\bar{T}_{\bullet} = \frac{\sum_{i=1}^k w_i T_i}{\sum_{i=1}^k w_i}$$

The weights that give the most precise estimate are

$$w_i = 1/v_i = 1/SE^2(T_i)$$

FIXED EFFECTS PROCEDURES

The variance (square of the standard error) of the weighted mean is the reciprocal of the sum of the weights

$$SE^2(\bar{T}_{\bullet}) = \frac{1}{\sum w_i}$$

A 95% confidence interval for the mean θ_{\bullet} is

$$\bar{T}_{\bullet} - 1.96SE \leq \theta_{\bullet} \leq \bar{T}_{\bullet} + 1.96SE$$

FIXED EFFECTS PROCEDURES

To test the hypothesis that the mean effect size parameter $\theta_{\bullet} = 0$

Use the statistic

$$Z = \bar{T}_{\bullet} / SE$$

which has the standard normal distribution when the null hypothesis is true

EXAMPLE: TYPE A BEHAVIOR AND CHD

Study	T	SE
1	-0.083	0.202
2	0.645	0.812
3	0.812	0.316
4	0.794	0.179
5	0.837	0.228

EXAMPLE: FIXED EFFECTS ANALYSIS OF TYPE A BEHAVIOR AND CHD

The *unweighted* average of the effect size estimates is
0.60

The *weighted* average of the effect size estimates is $\bar{T}_\bullet =$
0.555 with a standard error of
 $SE = 0.108$

The 95% confidence interval for the mean is
 $0.34 \leq \theta_\bullet \leq 0.77$

The z test statistic is $Z = 0.555/0.108 = 5.16, p < .01$

EXAMINING HETEROGENEITY

Do the studies all give the same answer?

Do all have the same effect size parameter?

We can answer this in two ways:

- A heterogeneity test
- An estimate of the variance of the effect size parameters across studies

EXAMINING HETEROGENEITY

A Heterogeneity Test

Use the test statistic

$$Q = \sum_{i=1}^k w_i (T_i - \bar{T}_{\bullet})^2$$

Q has a chi-square distribution with $(k - 1)$ df when

$$\theta_1 = \theta_2 = \dots = \theta_k$$

Reject homogeneity if Q is large (e.g., larger than the critical value of the chi-square)

EXAMINING HETEROGENEITY

Estimating the variance τ^2 of $\theta_1, \theta_2, \dots, \theta_k$

We do not observe $\theta_1, \theta_2, \dots, \theta_k$

To estimate τ^2 , we compute the excess variation in the effect size estimates beyond that expected by sampling error

The Q statistic measures variation and has expected value $(k - 1)$ when $\tau^2 = 0$.

EXAMINING HETEROGENEITY

Estimating the variance τ^2 of $\theta_1, \theta_2, \dots, \theta_k$

The estimate of τ^2 is

$$\hat{\tau}^2 = [Q - (k - 1)] / c$$

where

$$c = \left(\sum_{i=1}^k w_i - \frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i} \right)$$

whenever the estimate is greater than 0

EXAMINING HETEROGENEITY

Estimating the variance τ^2 of $\theta_1, \theta_2, \dots, \theta_k$

A useful statistic for describing the variation in the effect size parameters is $\hat{\tau}$, the square root of the variance component estimate—it represents the standard deviation of the effect parameters in the sample of studies observed

Comparing $\hat{\tau}^2$ to the average of the sampling error variances $(v_1 + v_2 + \dots + v_k)/k$ gives an indication of whether between-study variation exceeds within study (sampling error) variation

EXAMPLE: HETEROGENEITY ANALYSIS OF THE TYPE A BEHAVIOR DATA

The heterogeneity test statistic is

$$Q = 13.91, \text{ with } 5 - 1 = 4 \text{ df, } p < .001$$

The variance component estimate is

$$\hat{\tau} = (13.91 - 4)/62.73 = 0.158$$

so the estimate of τ is 0.398

The average v_i is 0.180, thus τ^2 is 88% as large

RANDOM EFFECTS PROCEDURES

In *random effects procedures*:

the goal is to estimate the mean population effect size μ_{θ} of the population from which the observed studies are a sample

between-study variation in effect sizes has an impact on weights and the uncertainty of this mean

RANDOM EFFECTS PROCEDURES

Between-study variation in effect size parameters is defined by the between studies variance component τ^2

The variance component is the variance of the effect size parameters in the population of studies from which the observed studies are a sample

RANDOM EFFECTS PROCEDURES

In the random effects model, the variance of the effect size estimates has two components:

- Within-study sampling error or conditional variance (measured by ν_i)
- Between-study variance in the effect parameters (measured by τ^2)

Thus the total variance of T_i is

$$\nu_i^* = \nu_i + \tau^2$$

RANDOM EFFECTS PROCEDURES

This suggests we should use a weighted mean like

$$\bar{T}_{\bullet}^* = \frac{\sum w_i^* T_i}{\sum w_i^*}$$

to estimate the average effect size.

The most precise estimate is given with weights

$$w_i^* = 1/v_i^* = 1/[SE^2(T_i) + \hat{\tau}^2]$$

RANDOM EFFECTS PROCEDURES

The variance (square of the standard error) of the weighted mean is the reciprocal of the sum of the weights

$$SE^2(\bar{T}_{\bullet}^*) = \frac{1}{\sum w_i^*}$$

A 95% confidence interval for the mean μ_{θ} is

$$\bar{T}_{\bullet}^* - 1.96SE \leq \mu_{\theta} \leq \bar{T}_{\bullet}^* + 1.96SE$$

RANDOM EFFECTS PROCEDURES

To test the hypothesis that the mean effect size parameter $\mu_{\theta} = 0$

Use the statistic

$$\mathbf{Z}^* = \bar{\mathbf{T}}_{\bullet}^* / \mathbf{SE}$$

which has the standard normal distribution when the null hypothesis is true

EXAMPLE: RANDOM EFFECTS ANALYSIS OF THE TYPE A BEHAVIOR DATA

The *unweighted* average of the effect size estimates is
0.60

The *weighted* average of the effect size estimates is $\bar{T}_{\bullet}^* =$
0.580 with a standard error of
 $SE = 0.223$

The 95% confidence interval for the mean is
 $0.14 \leq \mu_{\theta} \leq 1.02$

The z test statistic is $Z^* = .580/.223 = 2.60, p < .01$

EXAMPLE: COMPARISON OF FIXED AND RANDOM EFFECTS ANALYSES

	<u>Mean</u>	<u>SE</u>	<u>95% C. I.</u>
Fixed	0.555	0.108	0.34 to 0.77
Random	0.580	0.223	0.14 to 1.02

COMPARING FIXED AND RANDOM EFFECTS PROCEDURES

Random effects variances (v_i^*) are always larger than or equal to the fixed effects variances (v_i)-- they are equal only when $\tau^2 = 0$

This has several consequences:

- weights are smaller in random effects analyses than in fixed effects analyses
- weights are *more equal* in random effects analyses than in fixed effects analyses

COMPARING FIXED AND RANDOM EFFECTS PROCEDURES

- Standard errors of random effects analyses are usually larger than those of fixed effects analyses
- Random effects analyses are *more conservative*
- Random effects analyses are closer to unweighted analyses than fixed effects analyses

INFERENCES FROM SYSTEMATIC REVIEWS

Systematic reviews yield two kinds of inferences

- Inferences based on *aggregating* study effects (study generated evidence)
- Inferences based on *comparing* effects of different studies

The former are suitable for strong causal inference

The latter have the same status as observational studies