

Effect Sizes (ES) for Meta-Analyses

- ES – d, r/eta & OR
 - computing ESs
 - estimating ESs
 - ESs to beware!
- interpreting ES
- ES transformations
- ES adjustments
- outlier identification

Kinds of Effect Sizes

The effect size (ES) is the DV in the meta analysis.

d - standardized mean difference

- quantitative DV
- between groups designs

standardized gain score – pre-post differences

- quantitative DV
- within-groups design

r – correlation/eta

- converted from sig test (e.g., F, t, X^2) or set of means/stds
- between or within-groups designs or tests of association

odds ratio

- binary DVs
- between groups designs

Univariate (proportion or mean)

- prevalence rates

A useful ES:

- is standardized
- a standard error can be calculated

The Standardized Mean Difference (d)

- A Z-like summary statistic that tells the size of the difference between the means of the two groups
- Expresses the mean difference in Standard Deviation units
 - d = 1.00 → Tx mean is 1 std larger than Cx mean
 - d = .50 → Tx mean is 1/2 std larger than Cx mean
 - d = -.33 → Tx mean is 1/3 std smaller than Cx mean
- Null effect = 0.00
- Range from $-\infty$ to ∞
- Cohen's effect size categories
 - small = 0.20 medium = 0.50 large = 0.80

The Standardized Mean Difference (d)

$$\overline{ES} = \frac{\bar{X}_{G1} - \bar{X}_{G2}}{s_{pooled}}$$

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

- Represents a standardized group mean difference on an *inherently continuous* (quantitative) DV.
- Uses the pooled standard deviation
- There is a wide variety of d-like ESs – not all are equivalent
 - Some intended as sample descriptions while some intended as population estimates
 - define and use “n,” “n_k” or “N” in different ways
 - compute the variability of mean difference differently
 - correct for various potential biases

Equivalent formulas to calculate The Standardized Mean Difference (d)

- Calculate S_{pooled} using MS_{error} from a 2BG ANOVA

$$\sqrt{MS_{error}} = S_{pooled}$$

- Calculate S_{pooled} from F, condition means & ns

$$MS_{between} = \frac{\sum \bar{X}_j^2 n_j - \frac{(\sum \bar{X}_j n_j)^2}{\sum n_j}}{k - 1}$$

$$S_{pooled} = \sqrt{\frac{MS_{between}}{F}}$$

Equivalent formulas to calculate The Standardized Mean Difference (d)

- Calculate d directly from significance tests – t or F

$$ES = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

$$ES = \sqrt{\frac{F(n_1 + n_2)}{n_1 n_2}}$$

- Calculate t or F from exact p-value & df. Then apply above formulas.

$$ES = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

For t → <http://www.danielsoper.com/statcalc3/calc.aspx?id=10>

For F → <http://www.danielsoper.com/statcalc3/calc.aspx?id=7>



ds to beware!!!

- if you can get a mean difference & an error term, you can calculate d!!
- be careful where you get your mean differences !!
- you can use these, but carefully code what they represent!!!

- Corrected/estimated mean difference from ANCOVA
- b representing group mean comparison from a multivariate model

Both of these represent the part of the IV-DV effect that is independent of (controlling for) the other variables in the model

- This is a different thing than the bivariate IV-DV relationship!!!
- Be sure to code the specific variables being “controlled for” and the operationalization of the IV

ds to beware!!!

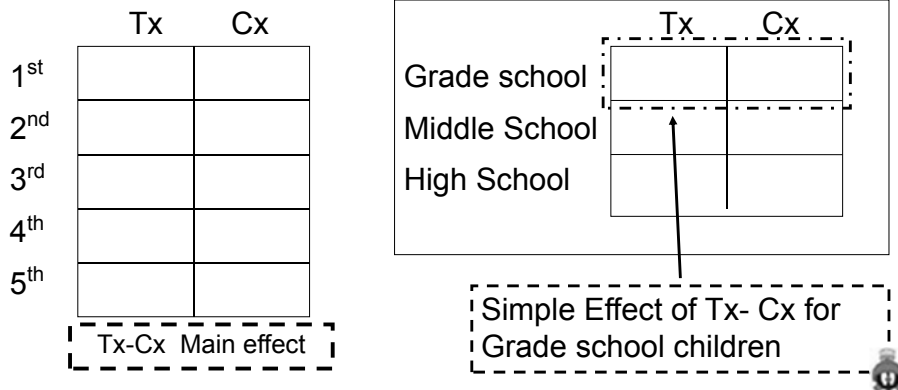
- if you can get a t or an F you can calculate d
- be careful where you get your ts & Fs !!
- you can use these, but carefully code what they represent!!!

- d calculated from t obtained from a multiple regression model...
- represents “unique” relationship between that variable and the criterion variable, after “controlling for” all the other variables in the model
 - only makes sense if the variable has 2 groups!!!
 - be sure to carefully code for what other variables are in the model & are being controlled for!

- d calculated from F obtained from ANCOVA or factorial ANOVA
- represents “unique” relationship between that variable and the criterion variable, after “controlling for” all the other variables in the model
 - only makes sense if the variable has 2 groups!!!
 - be sure to carefully code for what other variables are in the model & are being controlled for!


Getting the right effect size from a factorial design !!!

For example, you are conducting a meta analysis to estimate the effect size for comparisons of Tx & Cx among school children. You find the following studies – what means do you want to compare???



The Standardized Gain Score

- Like d , this is a Z-like summary statistic that tells the size of the difference between the means of the two groups
 - The “catch” is that there are three approaches to calculating it... (whichever you use → be sure to code BG v WG designs)
- Using the same S_{pooled} as d
 - Logic is that means and stds are same whether BG or WG, so d should be calculated the same
 - Using $\sqrt{MS_{error}}$ as S_{pooled}
 - Logic is that S_{pooled} should be based on “error variance” with subject variability excluded
 - Usually leads to larger effects sizes from WG designs than BG designs, even when both have same mean difference
 - Computing S_{pooled} using formula below
 - Similar logic to “2”, but uses a different estimate of S_{pooled}
 - S is the std of the gain scores
 - r is correlation between the pretest and posttest scores

$$S_{pooled} = \frac{S_{gain}}{\sqrt{2(1-r)}}$$


r / eta as “strength of effect” Effect Size

The advantage of r is that it can be used to include, in a single meta analysis, results from...

$$\text{BG or WG } t \quad ES = \sqrt{(t^2 / (t^2 + df))}$$

$$\text{BG or WG } F \quad ES = \sqrt{(F / (F + df))}$$

$$X^2 \quad ES = \sqrt{(X^2 / N)}$$

$$\text{Correlation} \quad ES = r$$

Also, r can be estimated whenever you have d

$$r = \sqrt{(d^2 / (4 + d^2))}$$

r “vs” eta....

You might see any of the formulas on the last page called “ r ” or “eta” – why both???

- r – is Pearson’s correlation – direction and strength of the linear relationship between the quantitative variables
- η - Eta – direction and strength of the relationship between the variables (linear and nonlinear) – must be positive!

They two converge for a 2-group design, but not for a k -group design, where the relationship between the group variable and the quantitative DV might be ...

- linear if grouping variable is quantitative (# practices)
- and/or nonlinear if grouping variable is quantitative
- an “aggregative of pairwise effect sizes” if grouping variable is qualitative



rs & etas to beware!!!

You can use them, but carefully code what they represent!!!

r/η calculated from F of a k-group designs

- can only be compared with η values from designs with “exactly the same” k groups
- be sure to code the specifics of the group operationalizations

partial η -- calculated by many statistical packages...

- calculated for multiple regression, GLM, ANCOVA, factorial ANOVA designs
- represent “unique” relationship between that variable and the criterion variable, after “controlling for” all the other variables in the model
- be sure to code for the specific variables that were controlled

rs & etas to beware!!!

You can use them, but carefully code what they represent!!!

partial & multiple partial correlations

- the correlation between two variables controlling both of them for one or multiple other variables
- be sure to code the specific variables that were controlled for

semi-partial & multiple semi-partial correlations

- the correlation between two variables controlling one of them for one or multiple other variables
- be sure to code for which variable is being controlled
- be sure to code the specific variables that were controlled for



Other Kinds of Correlations – can be used as ESs !!

Your friend & mine – Pearson’s Product-Moment Correlation

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Some of the usual formulas...

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

$$r = \frac{\sum z_x z_y}{N}$$

There are 2 other “kinds” of correlation:

- Computational short-cuts
 - applied when 1 or both variables are binary
 - produces the same Pearson’s r-value as the above formulas, but have fewer computational steps
- Estimation formulas
 - applied when 1 or both variables are binary
 - Estimate what Pearson’s would be if both variables were quantitative

Point-biserial Correlation

- pre-dates high-speed computers... calculators even...
- is a computational short cut that is applied when one variable is quantitative (ND) and the other is binary
- was very commonly used in test/scale development to compute item-total correlations
 - the correlation of each binary item with the total score computed from the sum of the items
 - “good items” were highly correlated with the total
- gives exactly the same value as the Pearson’s formulas!!
- only has full -1 to 1 range if binary variable is distributed as 50% / 50%!

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}} \quad \text{where...} \quad s_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Phi Correlation

- pre-dates high-speed computers, calculators even...
- is a computational short cut that is applied when both variables are binary
- was very commonly used in test/scale development to compute item-item correlations
 - the correlation of binary items with each other
 - “good items” were highly correlated with each other
- gives exactly the same value as the Pearson’s formulas!!
- only has full -1 to 1 range if both binary variables are distributed as 50% / 50%

	X^-	X^+	<i>Total</i>	
Y^-	a	b	e	$\phi = \frac{ad - bc}{\sqrt{efgh}}$
Y^+	c	d	f	
<i>Total</i>	g	h	n	

$$\Phi = \sqrt{(X^2 / N)}$$

Biserial Correlation

- is an estimation formula that is applied when
 - one variable is quantitative (ND) and the other is “quantitative but measured as binary”
 - you want to estimate what would Pearson’s correlation be if both had been measured as quantitative

$$r_b = (Y_1 - Y_0) \cdot (pq/Y) / \sigma_Y$$

Where...

- Y_1 & Y_0 are means of quantitative variable for each binary group
- p & q are proportions of sample in each binary group
- σ_Y is the population standard deviation of quantitative variable

There are further variations when one/both variables are rank-ordered.

Tetrachoric Correlation

- is an estimation formula that is applied when
 - both variables are “quantitative but measured as binary”
 - you want to estimate what would Pearson’s correlation be if both had been measured as quantitative

	X^-	X^+	Total
Y^-	a	b	e
Y^+	c	d	f
Total	g	h	n

$$r_{\text{tet}} = \cos (180/(1 + \sqrt{BC/AD})).$$

There are further variations when one/both variables are rank-ordered.



The Odds-Ratio

- Some meta analysts have pointed out that using the r-type or d-type effect size computed from a 2x2 table (binary DV & 2-group IV) can lead to an underestimate of the population effect size, to the extent that the marginal proportions vary from 50/50.
- A very workable alternative is to use the Odds-ratio !!!
- The odds-ratio is usually described as “the odds of success for Tx members, relative to the odds of success for Cx members.”
 - IV = Tx vs. Cx (coded 1 & 0)
 - DV = Improvement vs. No Improvement (coded 1 & 0)
 - Odds ratio of 2.5 means...
 - Those in the Tx group are 2.5 **times as likely** to show improvement as those in the Cx group

How to compute an odds-ratio

For these data*

IV male = 1 & female = 0

DV traditional = 1 & nontraditional = 0

GENDER * GROUP Crosstabulation

Count		GROUP		Total
		traditional	nontraditional	
GENDER	male	40	23	63
	female	102	123	225
Total		142	146	288

We are used to working with proportions

- the ratio of frequency in target category relative to total
- for males $40/63 \rightarrow .63492$ of males are traditional
- for females $102/225 \rightarrow .45333$ of females are traditional

Odds are computed differently:

- ratio of freq in target category relative to freq in other category
- males $40/23 \rightarrow 1.73913 \rightarrow$ if you are male, the odds are 1.73 to 1 that you are traditional
- females $102/123 \rightarrow .82927 \rightarrow$ if you are female, the odds are .83 to 1 that you are traditional

* Higher valued group coded as the comparison condition – coded = 0

How to compute an odds-ratio

For these data

IV male = 0 & female = 1

DV traditional = 0 & nontraditional = 1

GENDER * GROUP Crosstabulation

Count		GROUP		
		traditional	nontraditional	Total
GENDER	male	40	23	63
	female	102	123	225
Total		142	146	288

So, the odds-ratio is...

the odds ratio = $\frac{\text{the odds of being traditional for men}}{\text{odds of being traditional for women}}$

$$\text{odds ratio} = \frac{1.73913}{.82927} = 2.0972$$

Meaning → Males are 2.0972 times as likely to be traditional as women.

Computing the Odds-Ratio

The odds-ratio can be calculated directly from the frequencies of a 2x2 contingency table.

	Frequencies	
	Success	Failure
Treatment Group	<i>a</i>	<i>b</i>
Control Group	<i>c</i>	<i>d</i>

$$\overline{ES} = \frac{ad}{bc}$$

GENDER * GROUP Crosstabulation

Count		GROUP		
		traditional	nontraditional	Total
GENDER	male	40	23	63
	female	102	123	225
Total		142	146	288

$$ES = \frac{40 * 123}{23 * 102} = \frac{4920}{2346} = 2.0972$$

OR of 1 means no relationship between group & outcome

OR between 0 & 1 means a negative relationship

OR between 1 & infinity means a positive relationship

Considering Odds-Ratios

You need to be careful when considering odds-ratios !!!

Beware interpreting large, impressive looking, odds-ratios without checking the odds that are being "ratio-ed"!!!

	Succeed	Fail	
Tx	8	100000	ES = $\frac{800,000}{200,000} = 4.0$
Cx	2	100000	

Those who take the Tx are 4 times as likely to succeed as those who do not!!!

But check the odds for each... Tx 8/100000 = .00008

Cx 2/100000 = .00002

Not good odds in either group...!!!



Interpreting Effect Size Results

- Cohen's "Rules-of-Thumb"

- d

- small = 0.20
 - medium = 0.50
 - large = 0.80

- r

- small = 0.10
 - medium = 0.25
 - large = 0.40

- odds-ratio

- small = 1.50
 - medium = 2.50
 - large = 4.30

Rem – more important than these rules of thumb is knowing the "usual" effect sizes in your research area!

Wait! What happened to

.1, .3 & .5 for r ?????

Those translate to d-values of

.1, .62 & 1.15, respectively...

So, he changed them, a bit...

Also, these "adjusted" values better correspond to the distribution of effect sizes in published meta analyses as found by Lipsey & Wilson (1993)

Transformations

The most basic meta analysis is to take the average of the effect size from multiple studies as the best estimate of the effect size of the population of studies of that effect.

As you know, taking the average of a set of values "works better" if the values are normally distributed!

Beyond that, in order to ask if that mean effect size is different from 0, we'll have to compute a standard error of the estimated mean, and perform a Z-test. The common formulas for both of these also "work better" if the effect sizes are normally distributed.

And therein lies a problem! None of d, r & odds ratios are normally distributed!!!

So, it is a good idea to transform the data before performing these calculations !!

Transformations -- d

d has an upward bias when sample sizes are small

- the extent of bias depends upon sample size
- the result is that a set of d values (especially with different sample sizes) isn't normally distributed
- a correction for this upward bias & consequent non-normality is available

$$ES = d * \left[1 - \frac{3}{4N-9} \right]$$

Excel formula is $d * (1 - (3 / ((4*N) - 9)))$

Transformations -- r

r is not normally distributed

- and it has a problematic standard error formula.
- Fisher's Zr transformation is used – resulting in a set of ES values that are normally distributed

$$ES = .5 * \ln \left[\frac{1 + r}{1 - r} \right] \quad \text{Excel formula is } \text{FISHER}(r)$$

- all the calculations are then performed using the ES
- the final estimate of the population ES can be returned to r using another formula (don't forget this step!!!)

$$r = \frac{e^{2ES} - 1}{e^{2ES} + 1} \quad \text{Excel formula is } \text{FISHERINV}(ES)$$

Transformations – Odds-Ratio

the OR is asymmetrically distributed

- and has a complex standard error formula.
- one solution is to use the natural log of the OR
- nice consequence is that the transformed values are interpreted like d & r
 - Negative relationship < 0.
 - No relationship = 0.
 - Positive relationship > 0.

$$ES = \ln [OR] \quad \text{Excel formula is } \text{LN}(OR)$$

- all the calculations are then performed using the ES
- the final estimate of the population ES can be returned to OR using another formula (don't forget this step!!!)

$$OR = e^{ES} \quad \text{Excel formula is } \text{EXP}(ES)$$

Adjustments

(less universally accepted than transformations!!)

measurement unreliability

- what would r be if the DV were perfectly reliable?
- need reliability of DV (α)

$$r' = \frac{r}{\sqrt{\alpha_{DV}}}$$

range restriction

- What would r be if sample had full range of population DV scores ?
- "s" is sample std
- need unrestricted population std ("S")

Can use $r \leftrightarrow d$ formulas to obtain these

$$r' = \frac{S * r}{\sqrt{(S^2 r^2 + s^2 - s^2 r^2)}}$$

Adjustments, cont.
(less universally accepted than transformations!!)

artificial dichotomization of measures

–What would effect size be if variables had been measured as quantitative?

–If DV was dichotomized

- e.g., Tx-Cx & pass-fail instead of % correct
- use biserial correlation

–If both variables dichotomized

- e.g., some-none practices & pass-fail, instead of #practices & % correct
- Use tetrachoric correlation

Outlier Identification
(less universally accepted than transformations!!)

Outliers

- As in any aggregation, extreme values may have disproportionate influence
- Identification using Mosteller & Tukey method is fairly common
- Trimming and Winsorizing are both common

For all adjustments – Be sure to tell your readers what you did & the values you used for the adjustments!

