Institute for Health Research

Biostatistics Lunch
Lecture Series

Calculating and Reporting Effect Sizes

IHR BIOSTATISTICS LUNCH LECTURE SERIES PRESENTED BY
DR PAOLA CHIVERS
RESEARCH AND BIOSTATISTICS: INSTITUTE FOR HEALTH RESEARCH
THE UNIVERSITY OF NOTRE DAME AUSTRALIA
paola.chivers@nd.edu.au

Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude –not just, does a treatment affect people, but how much does it affect them.

-Gene V. Glass[1]

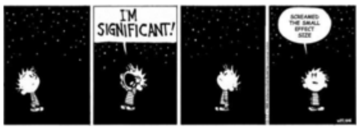
The primary product of a research inquiry is one or more measures of effect size, not P values.

-Jacob Cohen[2]

Quoted in Sullivan and Feinn (2012). Using effect size – or why the p value is not enough. Journal of Graduate Medical Education. September: 279-282. DOI: http://dx.doi.org/10.4300/JGME-D-12-00156.1

What are Effect Sizes?

A measurement of the size (magnitude) of an effect.
- independent of sample size
- standardized metric
- strength of association



Why are effect sizes important?

- Effect sizes help communicate practical importance of a result
  - an important but non significant result
  - an unimportant but significant result
- Provides information about 'how' important
- Allows comparisons across studies (meta analysis)
- Used to inform planning for future studies *a priori*
- Are a journal requirement

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

APA (7th ed.) guide to effect sizes

- Recommend inclusion for "readers to appreciate the magnitude or importance" (APA 2020 p. 89)
- Statistical estimate and should include confidence intervals

General Principal

"… provide the readers with enough information to assess the magnitude of the observed effect." (APA 2020 p. 89)

© 2020 Chivers IHR Biostatistics Lunch Lecture Series

Measuring Effect Size

Can be calculated in different ways.

Two main approaches:

1. Standardized difference between two means
   - Cohen's d
   - Hedge's d
   - Glass's Delta
   - Common language effect size (CLES)
2. Strength of association

Measuring Effect Size

1. Standardized difference between two means
   - *d* family
   - difference between observations divided by the standard deviation
   - Standard deviations units of effect
2. Strength of association
   - *r* family
   - proportion of variance that is explained by its group

Effect size measures:
Between subjects designs

Cohen's d (Cohen, 1988)

- Standardized mean difference of an effect
- Dependent variables can be measured on different scales or be completely different measurements (Lakens, 2013)
- Uncorrected effect size
  - Provides a biased estimate of the population effect size especially n<20

$$d_s = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}}$$

Fig. Cohen's d equation for the sample (Lakens, 2013 p3)

Effect size measures:
Between subjects designs

d calculated from the t-test differences between two groups

$$d_s = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Fig. Cohen's d equation for the sample related to a t-test (Lakens, 2013 p3)

Corrections for bias

Population effect size estimates based on sample averages overestimate the true population effect.

Corrections for bias can be applied:

- Corrections for Cohen's *d* = Hedges's *g*
- Corrections for eta squared (η^2) = omega squared (ω^2)

Effect size measures:
Between subjects designs

Hedges's d Standardized mean difference of an effect

- Corrected effect size
  - Provides an <u>unbiased</u> estimate of the population effect size

$$\text{Hedges's } g_s = \text{Cohen's } d_s \times \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right)$$

Fig. Hedges's g_s equation for the sample (Lakens, 2013 p3)

Effect size measures: Between subjects designs

Glass's Delta

- Used when standard deviations differ substantially between conditions.
- Choose the standard deviation of either the pre or post measurement.
  - Choose effect size that best represents the effect of interest.

Effect size measures: Two independent groups

Common language effect size (CLES)

- Non parametric effect size
- Probability of a z-score greater than a difference between groups of zero.
- Converts Cohen's d into a percentage, expressed as the probability
  - a random sampled person from one group is higher than a randomly sampled person from the other group (between design)
  - an individual has a higher value on one measurement compared to the other (within design)

See supplementary material from Lakens, 2013 for calculation spreadsheet. http://www.frontiersin.org/journal/10.3389/fpsyg.2013. 00863/abstract

Effect size measures: One sample or correlated data

- Uses the differences between measurements
- Standardized mean difference for within subjects design
- Cohen's d_z

$$\text{Cohen's } d_z = \frac{M_{\text{diff}}}{\sqrt{\frac{\sum (X_{\text{diff}} - M_{\text{diff}})^2}{N-1}}}$$

Fig. Cohen's d_z equation based on mean difference (Lakens, 2013 p4)

$$\text{Cohen's } d_z = \frac{t}{\sqrt{n}}$$

Fig. Cohen's d_z equation based on t-test (Lakens, 2013 p4)

Interpreting Effect Sizes

Interpreting Cohen's d

- Small *d*=0.2
- Medium *d*=0.5
- Large *d*=0.8

Note these are arbitrary
and the best way is to relate results to other effects reported in the literature.

What is the clinical importance of the result!

Reporting example

Independent t-test

An independent t-test did not report a significant difference between lower exam score for males (M=76.33 SD=5.84) compared to females (M=79.18 SD=6.89; t(32)=2.11 p=.071), although a medium to large effect was found (*d*=0.73 95% CI [.53-.86]).

** Note examples are fictitious with values presented for illustrative purposes only.*

Effect size measures:
Three or more independent groups (e.g. ANOVA)

- Eta squared η^2 (within)
- Partial Eta squared η_p^2 (between studies)

→ Degree of association for sample

- Omega squared
- Intraclass correlation

→ Degree of association for population

Interpreting Effect Sizes

Interpreting Eta squared

- Small *.01 or 1%*
- Medium *.06 or 6%*
- Large *.138 or 13.8%*

Pallant (2020) p 218

REMEMBER these are arbitrary and the best way is to relate results to other effects reported in the literature.

What is the clinical importance of the result!

Reporting example

Dependent t-test

An dependent t-test reported a significant increase in 3K time trial from pre season (M=189 sec SD=33) to mid season (M=151 sec SD=16; t(57)=6.91 p=.002), with a large effect (η^2=.37).

** Note examples are fictitious with values presented for illustrative purposes only.*

Reporting example

ANOVA

Group differences examined using ANOVA found independent schools had significantly higher ATAR scores (M=89.1 SD=5.2), compared to Catholic schools (M=82.3 SD=3.5) who were higher than public schools (M=73.7 SD=16.2; F(2, 1654)=4.53 p=.019). Despite reaching statistical significance, the actual mean difference between school types was quite small (η^2=.02).

** Note examples are fictitious with values presented for illustrative purposes only.*

Effect size measures: Chi-square tests

- Phi
  - Two binary variables
  - Related to correlation and Cohen's *d*
  - Interpreted like Pearson's r and R^2
- Cramer's Phi or V
  - More than two categorical variables
  - Measures inter-correlation
  - Biased as increases with the number of cells

Interpreting Effect Sizes

Interpreting phi (2x2)

- Small .10
- Medium *.30*
- Large *.50*

Interpreting Cramer's V

- Two categories:
  - Small .01, Medium *.30,* Large *.50*
- Three categories:
  - Small .07, Medium *.21,* Large *.35*
- Four categories:
  - Small .06, Medium *.17,* Large *.29*

Pallant (2020) p 228

REMEMBER these are arbitrary and the best way is to relate results to other effects reported in the literature.

What is the clinical importance of the result!

Reporting example

Chi square

A chi square group difference between sex and employment status indicated a significant association of medium effect (χ^2 (1, n=186)=12.56 p=.003, *phi*=.39). More females were employed part-time compared to males (53% versus 39%), while more males were in full time employment (61% versus 47%).

** Note examples are fictitious with values presented for illustrative purposes only.*

Calculating effect sizes

Abundance of online calculators

Lenhard and Lenhard (2016) provide a number of different online calculators based on your sample and study design.

Lenhard, W. & Lenhard, A. (2016). *Calculation of Effect Sizes*. Retrieved from: **https://www.psychometrica.de/effect_size.html**. Dettelbach (Germany): Psychometrica. DOI: 10.13140/RG.2.2.17823.92329

Reporting example – in text

Small-sided games can discriminate perceptual-cognitive-motor capability and predict disposal efficiency in match performance of skilled Australian footballers

Ben Piggott, Sean Müller, Paola Chivers, Ashley Cripps & Gerard Hoyne

Figure 1. Mann-Whitney U tests reported a significantly higher median score between higher (Decision: *Median* = 2.90, *Range* = 0.30; Total: *Median* = 5.36, *Range* = 1.13) and lower skilled (Decision: *Median* = 2.80, *Range* = 0.73; Total: *Median* = 5.11, *Range* = 1.97) players for decision (p = .012, d = 0.26) and total score (p = .037, d = 0.30). There was no significant difference between the two groups for execution score (p = 0.90), although higher skilled players still performed at a higher level (d = 0.25).

The LMM analysis found that the set the trial occurred in and trial number were not significant factors in the test total score model and were removed from the final reported model. The first LMM (Table 2) found higher skilled players scored significantly higher than lower skilled players on the total score (β = 0.45, p = .009). Kruskal-Wallis tests showed that there was a statistically significant difference in decision scores ($\chi^2 2$) = 8.82, p = 0.012, between players with 1–25 games (M = 2.86) and 51+ games (M = 2.98), as well as between players with 26–50 games (M = 2.79) and 51+ games (M = 2.98). Large effect sizes were found between the decision scores of 51+ and 26–50 (d = 2.01), as well as 51+ and 1–25 (d = 2.23) games played. There was no significant difference

Piggott et al., 2018 p 4

Reporting example - Tables

Manipulating Field Dimensions During Small-sided Games Impacts the Technical and Physical Profiles of Australian Footballers

Fleay, Brock, Joyce, Christopher, Banyard, Harry, and Woods, Carl T. (2018) Manipulating field dimensions during small-sided games impacts the technical and physical profiles of Australian footballers. *Journal of Strength and Conditioning Research*, 32 (7). pp. 2039-2044.

Table 3. Descriptive statistics (mean ± *SD*) and between-group effects (d) for each SSG.*

Criterion variable	Small	Medium	Large	Small-medium d	Medium-large d	Small-large d
Handball (n)	19.6 ± 3.8	19.9 ± 4.0	19.0 ± 4.2	−0.07 (trivial)	0.21 (moderate)	0.15 (moderate)
Tackle (n)†	5.6 ± 1.7	4.7 ± 2.4	4.1 ± 2.3	0.45 (moderate)	0.25 (moderate)	0.75 (large)
Turnover (n)†‡	3.9 ± 1.6	3.5 ± 1.2	2.4 ± 1.4	0.28 (moderate)	0.82 (very large)	0.99 (very large)
Effective handball (n)	16.4 ± 4.0	16.7 ± 4.5	16.9 ± 3.9	−0.07 (trivial)	−0.05 (trivial)	−0.14 (small)
Ineffective handball (n)†‡	3.3 ± 1.6	3.2 ± 1.7	2.1 ± 1.5	0.02 (trivial)	0.66 (large)	0.72 (large)
Effective tackle (n)	2.9 ± 1.4	2.7 ± 1.4	2.4 ± 1.6	0.16 (small)	0.24 (moderate)	0.39 (moderate)
Ineffective tackle (n)	2.6 ± 1.4	2.0 ± 1.5	2.0 ± 1.4	0.43 (moderate)	−0.05 (trivial)	0.39 (moderate)
Shepherd (n)	2.3 ± 1.8	2.1 ± 1.7	2.9 ± 1.7	0.07 (trivial)	−0.44 (moderate)	−0.35 (moderate)
Interception (n)	1.1 ± 1.4	1.0 ± 0.9	0.6 ± 0.8	0.07 (trivial)	0.63 (large)	0.53 (large)
Bounce (n)†	0.04 ± 0.2	0.1 ± 0.3	0.3 ± 0.5	−0.42 (moderate)	−0.37 (moderate)	−0.79 (large)
Goal	2.6 ± 1.3	2.1 ± 0.9	2.1 ± 0.9	0.36 (moderate)	0.09 (trivial)	0.43 (moderate)
Total distance (m)†‡§	153.7 ± 9.7	176.8 ± 18.8	195.9 ± 15.6	−1.53 (very large)	−1.10 (very large)	−3.22 (very large)
Relative distance ($m \cdot min^{-1}$)†‡§	150.9 ± 9.7	170.2 ± 16.4	184.0 ± 10.8	−1.44 (very large)	−1.00 (very large)	−3.23 (very large)
Maximum velocity ($km \cdot h^{-1}$)†‡§	19.2 ± 1.2	21.2 ± 1.0	22.1 ± 1.3	−1.78 (very large)	−0.78 (large)	−2.30 (very large)
PlayerLoad (AU)†‡§	19.5 ± 0.4	20.8 ± 0.4	22.1 ± 0.4	−3.15 (very large)	−2.99 (very large)	−6.21 (very large)
Slow jogging distance (m)†‡§	89.6 ± 3.4	82.8 ± 4.6	76.6 ± 4.2	1.67 (very large)	1.39 (very large)	3.39 (very large)
Fast jogging distance (m)†‡§	9.0 ± 2.9	13.9 ± 3.7	18.2 ± 3.2	−1.45 (very large)	−1.24 (very large)	−3.01 (very large)
Sprinting distance (m)†‡§	1.2 ± 0.9	3.2 ± 1.3	5.0 ± 2.0	−1.75 (very large)	−1.07 (very large)	−2.46 (very large)

*SSG = small-sided game.

Fleay et al., 2018 p 2042

Cautions when using Effect Sizes

- Reported cut-points for interpreting effect sizes are arbitrary and the best way is to relate results to other effects reported in the literature.
- Effect sizes are sensitive to spurious influences, such as:
  - which standard deviation is used, with pooled SD better
  - pooled SD based on assumption of estimates from the same population
  - whether there have been corrections for bias
  - whether data has a normal distribution
  - reliability of the measurement
- Care when comparing effects sizes of outcomes
- 'effect size' expression implies causality so should be used appropriately i.e. is this implication intended or justified.

Summary - effect sizes

- Effect sizes help communicate practical importance of a result
  - an important but non significant result
  - an unimportant but significant result
- Can be a journal requirement
- Recommended according to APA style
- General Principal "… provide the readers with enough information to assess the magnitude of the observed effect."
  (APA 2020 p. 89)

Summary Interpreting Effect Sizes

Table from Lenhard and Lenhard (2016)

d	r*	η^2	Interpretation sensu Cohen (1988)	Interpretation sensu Hattie (2009)
< 0	< 0	-	Adverse Effect	
0.0	.00	.000	No Effect	Developmental effects
0.1	.05	.003		
0.2	.10	.010	Small Effect	Teacher effects
0.3	.15	.022		
0.4	.2	.039		Zone of desired effects
0.5	.24	.060	Intermediate Effect	
0.6	.29	.083		
0.7	.33	.110		
0.8	.37	.140	Large Effect	
0.9	.41	.168		
≥ 1.0	.45	.200		

See Coe (2002) for BESD and CLES interpretation Table I.
https://www.leeds.ac.uk/educol/documents/00002182.htm

Summary Interpreting Effect Sizes

Index	Description[b]	Effect Size	Comments
Between groups			
Cohen's d[a]	$d = M_1 - M_2 / s$ $M_1 - M_2$ is the difference between the group means (M); s is the standard deviation of either group	Small 0.2 Medium 0.5 Large 0.8 Very large 1.3	Can be used at planning stage to find the sample size required for sufficient power for your study
Odds ratio (OR)	$\frac{\text{Group 1 odds of outcome}}{\text{Group 2 odds of outcome}}$ If OR = 1, the odds of outcome are equally likely in both groups	Small 1.5 Medium 2 Large 3	For binary outcome variables Compares odds of outcome occurring from one intervention vs another
Relative risk or risk ratio (RR)	Ratio of probability of outcome in group 1 vs group 2; If RR = 1, the outcome is equally probable in both groups	Small 2 Medium 3 Large 4	Compares probabilities of outcome occurring from one intervention to another
Measures of association			
Pearson's r correlation	Range, −1 to 1	Small ±0.2 Medium ±0.5 Large ± 0.8	Measures the degree of linear relationship between two quantitative variables
r^2 coefficient of determination	Range, 0 to 1; Usually expressed as percent	Small 0.04 Medium 0.25 Large 0.64	Proportion of variance in one variable explained by the other

Table 1 from Sullivan and Feinn (2012). Using effect size – or why the p value is not enough. Journal of Graduate Medical Education. September: 279-282. DOI: http://dx.doi.org/10.4300/JGME-D-12-00156.1

Summary of Effect Sizes *d* family

Table 1 from Lakens, 2013 p5

ES	Standardizer	Use
Cohen's d_{pop}	σ (population)	Independent groups, use in power analyses when population σ is known, σ calculated with n
Cohen's d_s	Pooled SD	Independent groups, use in power analyses when population σ is unknown, σ calculated with n-1
Hedges' g	Pooled SD	Independent groups, corrects for bias in small samples, report for use in meta-analyses
Glass's Δ	SD pre measurement or control condition	Independent groups, use when experimental manipulation might affect the SD
Hedges' g_{av}	$(SD_1 + SD_2)/2$	Correlated groups, report for use in meta-analyses (generally recommended over Hedges' g_{rm})
Hedges' g_{rm}	SD difference scores corrected for correlation	Correlated groups, report for use in meta-analyses (more conservative then Hedges' g_{av})
Cohen's d_z	SD difference scores	Correlated groups, use in power analyses

Summary of Effect Sizes *r* family

Table 2 from Lakens, 2013 p7

ES (Biased)	ES (Less Biased)	Use
eta squared (μ^2)	omega squared (ω^2)	Use for comparisons of effects within a single study
eta squared (μ_p^2)	omega squared (ω_p^2)	Use in power analyses, and for comparisons of effect sizes across studies with the same experimental design.
Generalized eta squared (μ_G^2)	Generalized omega squared (ω_G^2)	Use in meta-analyses to compare across experimental designs

References and further reading

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). Https://doi.org/10.1037/0000165-000

Coe, R. (2002). It's the effect size, stupid: What effect size is and why it is important. Paper presented at the Annual conference of the British Education Research Association, University of Exeter, England, 12-14 September. https://www.leeds.ac.uk/educol/documents/00002182.htm

Lakens D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Frontiers in Psychology 4; 836. https://www.frontiersin.org/article/10.3389/fpsyg.2013.00863

Lenhard, W. & Lenhard, A. (2016). Calculation of Effect Sizes. Retrieved from: https://www.psychometrica.de/effect_size.html. Dettelbach (Germany): Psychometrica. DOI: 10.13140/RG.2.2.17823.92329

Nandy, K. Understanding and quantifying effect sizes. https://www.sonnet.ucla.edu/sites/default/files/son_files/RESEARCH/effect_size_4-9-2012.pdf

Palant, J. 2020. SPSS survival manual: A step by step guide to data analysis using IBM SPSS (7th ed.). Allen & Unwin.

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. Journal of graduate medical education, 4(3), 279–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/