1   What measure of effect size when comparing two groups based on their means?

2   Marie Delacre[1] & Christophe Leys[1]

3   [1] Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, Belgium

4   Author Note

5   Correspondence concerning this article should be addressed to Marie Delacre, CP191,

6   avenue F.D. Roosevelt 50, 1050 Bruxelles. E-mail: marie.delacre@ulb.ac.be

Abstract

Keywords: keywords

Word count: X

¹¹   What measure of effect size when comparing two groups based on their means?

Intro

¹³   During decades, researchers in social science (Henson & Smith, 2000) and education

¹⁴ (Fan, 2001) have overestimated the ability of the null hypothesis (H0) testing to determine

¹⁵ the importance of their results. The standard for researchers in social science is to define H0

¹⁶ as the absence of effect (Meehl, 1990). For example, when comparing the mean of two

¹⁷ groups, researchers commonly test the H0 that there is no mean differences between groups

¹⁸ (Steyn, 2000). Any effect that is significantly different from zero will be seen as sole support

¹⁹ for a theory.

²⁰   Such an approach has faced many criticisms among which the most relevant to our

²¹ concern is that the null hypothesis testing highly depends on sample size: for a given alpha

²² level and a given difference between groups, the larger the sample size, the higher the

²³ probability of rejecting the null hypothesis (Fan, 2001; Kirk, 2009; Olejnik & Algina, 2000;

²⁴ Sullivan & Feinn, 2012). It implies that even tiny differences could be detected as

²⁵ statistically significant with very large sample sizes (McBride, Loftis, & Adkins, 1993)[1].

²⁶   Facing this argument, it has become an adviced practice to report the *p*-value assorted

²⁷ by a measure of the effect size, that is, a quantitative measure of the magnitude of the

²⁸ experimental effect (Cohen, 1965; Fan, 2001; Hays, 1963). This practice is also highly

²⁹ endorsed by the American Psychological Association (APA) and the American Educational

³⁰ Research Association (AERA) (American Educational Research Association, 2006; American

³¹ Psychological Association, 2010). However, limited studies properly report effect size in the

———

[1] Tiny differences might be due to sampling error, or to other factors than the one of interest: even under
the assumption of random assignent (which is a necessary but not sufficient condition), it is almost
impossible to be sure that the only difference between two conditions is the one defined by the factor of
interest. Other tiny factors of no theoretical interest might slighly influence results, making the probability of
getting an actual zero effect very low. This is what Meehl (1990) calls 'systematic noise'

32 last several decades.

33   First, there is a high confusion between the effect size and other related concepts such
34 as the clininal significance [SEE NOTE LATER] of a result (i.e. the relevance of an effect in
35 real life). Moreover, there are several situations that call for effect size measures and in the
36 current litterature, it's not always easy to know which measure using in a specific context.

37   Second, when used for inference, the main measures of effect sizes (i.e. Cohen's d and
38 point-biserial r) are submitted to a range of assumptions (i.e. normality and
39 heteroscedasticity) and these assumptions are known to be unrealistic in many research
40 designs (Cain, Zhang, & Yuan, 2017; Erceg-Hurn & Mirosevich, 2008; Glass, Peckham, &
41 Sanders, 1972; Grissom, 2000; Micceri, 1989; Yuan, Bentler, & Chan, 2004). As consequences
42 many estimations of effect size are inaccurate and alter the robustness of the statistical
43 conclusions. In the context of comparing two groups based on their means, Cohen's d_s is the
44 dominant effect size measure used by researchers (Peng, Chen, Chiang, & Chiang, 2013;
45 Shieh, 2013). We will argue that, like Student's t-test, this measure rely on the often
46 untenable assumptions of normality and homogeneity of variances (Cumming, 2013; Grissom
47 & Kim, 2005; Kelley, 2005; Shieh, 2013). While it is becoming more common in statistical
48 software to present Welch's t-test by default, when performing a t-test (i.e., R, Minitab),
49 similar issues for the measures of effect sizes has received less attention (Shieh, 2013) and
50 Cohen's d_s remains persistent [2].

51   In sum the aim of this paper is threefold:
52   1. Clearly define what is (and what is not) a measure of effect size;
53   2. Listing the different situations that call for effect sizes measure and reviewing which
54 measure is appropriate in which circumstance;
55   3. Define different properties of a good effect size estimator and discuss the impact of

―――――

[2] For example, in Jamovi, Cohen's ds is provided, whatever one performs Student's or Welch's t-test

56 assumptions violations on the robustness of the measures of effect size, based on simulations.

Measure of effect size: what it is, what it is not

58   The effect size is commonly refered to the practical significance of a test. Grissom and
59 Kim (2005) define the effect size as the extent to which results differ from what is implied by
60 the null hypothesis. In the context of the comparison of two groups based on their mean,
61 depending on the defined null hypothesis (considering the absence of effect as the null
62 hypothesis), we could define the effect size either as the magnitude of differences between
63 parameters of two populations groups are extracted from (e.g. the mean; Peng & Chen, 2014)
64 or as the magnitude of the relation between one dichotomous factor and one dependent
65 variable (American Educational Research Association, 2006). Both definitions refers to as
66 the most famous families of measures of effect sizes [Rosenthal_1994]: respectively the
67 d-family and the r-family.

68   Very often, the contribution of the measures of effect size is overestimated.

69   First, benchmarks about what should be a small, medium or large effect size might
70 have contributed at seeing the effect size as a measure of the importance or the relevance of
71 an effect in real life, but it is not (Stout & Ruble, 1995). The effect size is only a
72 mathematical indicator of the magnitude of a difference, which depends on the way a
73 variable is converted into numerical indicator. In order to assess the meaningfulness of an
74 effect, we should be able to relate this effect with behaviors/meaningful consequences in the
75 real world (Andersen, McCullagh, & Wilson, 2007). For example, let us imagine a sample of
76 students in serious school failure who are randomly divided into two groups: an experimental
77 group following a training program and a control group. At the end of the training, students
78 in the experimental group have on average significantly higher scores on a test than students
79 in the control group, and the difference is large (e.g. 30 percents). Does it mean that
80 students in the experimental condition will be able to pass to the next grade and to continue

normal schooling? Whether the computed magnitude of difference is an important,
meaningful change in everyday life refers to another construct: the *clinical significance*
(Bothe & Richardson, 2011). [I DON'T LIKE THE WORD "CLINICAL" BECAUSE IT'S
NOT GENERAL ENOUGH.A MEANINGFUL SIGNIFICANCE COULD BE SOCIAL,
PERSONAL, CLINICAL, PROFESIONNAL... ANY IDEA OF A MORE GENERAL
WORD?]. It refers to the interpretation of treatment outcomes and is neither statistical nor
mathematical, it is related to underlying theory that posits an empirical hypothesis. In other
words, the relation between *practical* and *clinical* significance is more a theoretical argument
than a statistical one.

Second, in the context of the comparison of two groups based on their means, it should
not replace the null hypothesis testing. Statistical testing allows the researcher to determine
whether the oberved departure from H0 occured by chance or not (Stout & Ruble, 1995)
while effect size estimators allow to assess the practical signficance of an effect, and as
reminds Fan (2001) "a practically meaningful outcome may also have occured by chance, and
consequently, is not trustworthy". For this reason, the use of confidence intervals around the
effect size estimate is highly recommended (Bothe & Richardson, 2011).

Different goals of measures of effect sizes

Effect size measures can be used for *inferential* purposes:
- The effect sizes from previous studies can be used in a priori power analysis when
planning a new study (Lakens, 2013; Prentice & Miller, 1990; Stout & Ruble, 1995; Sullivan
& Feinn, 2012; Wilkinson & the Task Force on Statistical Inference, 1999)
- To compute confidence intervals (Shieh, 2013)

Measures of effect size can also be used for *comparative* purposes, i.e. to assess the
stability of results across designs, analysis, samples sizes (Wilkinson & the Task Force on
Statistical Inference, 1999). It includes:

106   - To compare results of 2 or more studies (Prentice & Miller, 1990)

107   - To incorporate results in meta-analysis (Lakens, 2013; Li, 2016; Nakagawa & Cuthill,

108 2007; Stout & Ruble, 1995; Wilkinson & the Task Force on Statistical Inference, 1999)

109   Finally, effect size measures can be used for *interpretive* purposes: in order to assess

110 the practical significance of a result (beyond statistical significance; Lakens, 2013; American

111 Psychological Association, 2010; Prentice & Miller, 1990)

112   When used for *inference*, effect size measures should meet some mathematical

113 properties (see next section). According to the statistical properties of Welch's statistic, in

114 the context of heteroscedasticity, it does not seem possible to define a proper measure of

115 effect size without taking the specificity of the design (and more specifically the sample size

116 allocation ratio) into account in the calculation of the group variance (Shieh, 2013). On the

117 other side, generality is required for comparative and interpretive purposes (Cumming, 2013).

Robust measures

Properties of a good effect size estimator

120   **Unbiasedness.**

121   **Consistency.**

122   **Efficiency.**

123   **Interpretability.** Interpretability is not a very clear concept to me. Reading about

124 it!

Simulations

126   American Educational Research Association. (2006). Standards for reporting on

127 empirical social science research in aera publications. *Educational Researcher*, *35*, 33–40.

128 doi:10.3102/0013189X035006033

129   American Psychological Association. (2010). *Publication manual of the american*

psychological association [apa] (6 ed.) (American Psychological Association.). Washington, DC:

Andersen, M. B., McCullagh, P., & Wilson, G. J. (2007). But what do the numbers really tell us? Arbitrary metrics and effect size reporting in sport psychology research. *Journal of Sport & Exercise Psychology*, *29*, 664–672.

Bothe, A. K., & Richardson, J. D. (2011). Statistical, practical, clinical, and personal significance: Definitions and applications in speech-language pathology. *American Journal of Speech-Language Pathology*, *20*, 233–242.

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. doi:10.3758/s13428-016-0814-1

Cohen, J. (1965). Some statistical issues in psychological research. In *Handbook of clinical psychology* (B. B. Wolman., pp. 95–121). New York: McGraw-Hill.

Cumming, G. (2013). Cohen's d needs to be readily interpretable: Comment on shieh (2013). *Behavior Research Methods*, *45*, 968–971. doi:10.3758/s13428-013-0392-4

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*(7), 591–601. doi:10.1037/0003-066X.63.7.591

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research*, *94*(5), 275–282. doi:10.1080/00220670109598763

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. doi:10.3102/00346543042003237

153   Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting*
154 *and Clinical Psychology*, *68*(1), 155–165. doi:10.1037//0022-006x.68.1.155

155   Grissom, R. R., & Kim, J. J. (2005). *Effect size for research: A broad practical*
156 *approach.* (Lawrence Erlbaum Associates, Mahwah, N.J.). London.

157   Hays, W. L. (1963). *Statistics for psychologists* (Holt, Rinehart & Winston.). New
158 York.

159   Henson, R. I., & Smith, A. D. (2000). State of the art in statistical significance and
160 effect size reporting: A review of the APA task force report and current trends. *Journal of*
161 *Research and Development in Education*, *33*(4), 285–296.

162   Kelley, K. (2005). The effects of nonnormal distributions on confidence intervales
163 around the standardized mean difference: Bootstrap and parametric confidence intervals.
164 *Educational and Psychological Measurement*, *65*(1), 51–69. doi:10.1177/0013164404264850

165   Kirk, R. E. (2009). Practical significance: A concept whose time has come. *Educational*
166 *and Psychological Measurement*, *56*(5), 746–759. doi:10.1177/0013164496056005002

167   Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative
168 science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(863), 1–12.
169 doi:10.3389/fpsyg.2013.00863

170   Li, J. (2016). Effect size measures in a two-independent-samples case with nonnormal
171 and nonhomogeneous data. *Behavior Research Methods*, *48*(4), 1560–1574.
172 doi:10.3758/s13428-015-0667-z

173   McBride, G. B., Loftis, J. C., & Adkins, N. C. (1993). What do significance tests really
174 tell us about the environment? *Environmental Management*, *17*(4), 423–432.

175   Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian

defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108–141.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166. doi:10.1037/0033-2909.105.1.156

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews, 82*, 591–605. doi:10.1111/j.1469-185X.2007.00027.x

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241–286. doi:10.1006/ceps.2000.1040

Peng, C.-Y., & Chen, L.-T. (2014). Beyond cohen's d: Alternative effect size measures for between-subject designs. *THE JOURNAL OF EXPERIMENTAL EDUCATION, 82*(1), 22–50. doi:10.1080/00220973.2012.745471

Peng, C.-Y., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The Impact of APA and AERA Guidelines on Effect size Reporting. *Contemporary Educational Psychology, 82*(1), 22–50. doi:10.1080/00220973.2012.745471

Prentice, D., & Miller, D. T. (1990). When small effects are impressive. *Psychological Bulletin, 112*(1), 160–164.

Shieh, G. (2013). Confidence intervals and sample size calculations for the weighted eta-squared effect sizes in one-way heteroscedastic ANOVA. *Behavior Research Methods, 45*(1), 2–37. doi:10.3758/s13428-012-0228-7

Steyn, H. S. (2000). Practical significance of the difference in means. *Journal of Industrial Psychology, 26*(3), 1–3.

Stout, D. D., & Ruble, T. L. (1995). Assessing the practical signficance of empirical

results in accounting education research: The use of effect size information. *Journal of Accounting Education*, *13*(3), 281–298.

Sullivan, G., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, 279–282. doi:10.4300/JGME-D-12-00156.1

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. doi:10.1007/bf02295644