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 Generally, there is a high confusion between the effect size and other related concepts

34 such as the Applied significance[2]. Moreover, there are several situations that call for effect

35 size measures and in the current litterature, it's not always easy to know which measure using

36 in a specific context. We will therefore introduce this paper with 3 sections in which we will:

37 1. Clearly define what is a measure of effect size;

38 2. Listing the different situations that call for effect sizes measure;

39 3. Define required properties as a function of the situations.

40 After these general adjustments, we will focus our attention on "between-subject"

41 designs where individuals are randomly assigned into one of two independant groups and

42 groups scores are compared based on their means[3]. Because it has been widely argued that

43 there are many fields in psychology where the assumption of equal variances between groups

44 is ecologically unlikely (Erceg-Hurn & Mirosevich, 2008; Grissom, 2000), it is becoming more

45 common in statistical software to present a t-test that does not hold on this assumption by

46 default, namely the Welch's t-test (e.g., R, Minitab). However, similar issues for the

47 measures of effect sizes has received less attention (Shieh, 2013), and Cohen's d_s remains

48 persistent[4]. One possible reason is that researchers cannot find a consensus on which

49 alternative should be in use (Shieh, 2013). We will limit our study to the standardized mean

50 difference, called the d-family, because it is the dominant family of estimators of effect size

51 when comparing two groups based on their means (Peng, Chen, Chiang, & Chiang, 2013;

52 Shieh, 2013), and we will see that even in this very specific context, there is little agreement

53 between researchers as to which is the most suitable estimator. According to us, the main

---

[2] In our conception Applied significance" encompass all what refers to the relevance of an effect in real life, e.g. clinical, personnal, social, professionnal.

[3] We made this choice because *t*-tests are still the most commonly used tests in the field of Psychology.

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

54 reason is that it is difficult, based on currently existing measures, to optimally serves all the

55 purposes of en affect size measure. Throughout this section, we will:

56   1. Present the main measures of the *d*-family that are proposed in the

57 literature,related to the purpose they serve, and introduce a new one, namely the

58 "transformed Shieh's *d*" that should help at reaching all the purposes simultaneously;

59   2. Present and discuss the results of simulations we performed, in order to compare

60 existing measures and the new introduced one;

61   3. Summarize our conclusions in practical recommandations.

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

63   The effect size is commonly refered to the practical significance of a test. Grissom &

64 Kim (2005) define the effect size as the extent to which results differ from what is implied by

65 the null hypothesis. In the context of the comparison of two groups based on their mean,

66 depending on the defined null hypothesis (considering the absence of effect as the null

67 hypothesis), we could define the effect size either as the magnitude of differences between

68 parameters of two populations groups are extracted from (e.g. the mean; Peng & Chen, 2014)

69 or as the magnitude of the relation between one dichotomous factor and one dependent

70 variable (American Educational Research Association, 2006). Both definitions refers to as

71 the most famous families of measures of effect sizes [Rosenthal_1994]: respectively the

72 *d*-family and the *r*-family.

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

74 benchmarks about what should be a small, medium or large effect size might have

75 contributed at seeing the effect size as a measure of the importance or the relevance of an

76 effect in real life, but it is not (Stout & Ruble, 1995). The effect size is only a mathematical

77 indicator of the magnitude of a difference, which depends on the way a variable is converted

78 into numerical indicator. In order to assess the meaningfulness of an effect, we should be

79 able to relate this effect with behaviors/meaningful consequences in the real world

(Andersen, McCullagh, & Wilson, 2007). For example, let us imagine a sample of students in serious school failure who are randomly divided into two groups: an experimental group following a training program and a control group. At the end of the training, students in the experimental group have on average significantly higher scores on a test than students in the control group, and the difference is large (e.g. 30 percents). Does it mean that 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 *Applied significance* (Bothe & Richardson, 2011). 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 *Applied* 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"* (p.278). For this reason, the use of confidence intervals around the effect size estimate is highly recommended (Bothe & Richardson, 2011).

Different purposes of effect size measures

Effect size measures can be used in an *inferential* perspective:

- 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)

105   When related to inference, effect size measures are generally submitted to a range of

106 assumptions (e.g. independent and identically distributed residuals are normal and have

107 equal variances between groups). When these assumptions are not met, many estimations of

108 effect size are inaccurate and alter the robustness of the statistical conclusions.

109   Measures of effect size can also be used in a *comparative* perspective, that is to assess

110 the stability of results across designs, analysis, samples sizes (Wilkinson & the Task Force on

111 Statistical Inference, 1999). It includes:

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

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

114 2007; Stout & Ruble, 1995; Wilkinson & the Task Force on Statistical Inference, 1999)

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

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

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

118   **Properties of a good effect size estimator**

119   The estimate of an estimator depends on the sampling. That is to say, based on

120 different samples extracted from the same population, one would obtain different estimates

121 of the same estimator. The *sampling distribution* of the estimator is the distribution of all

122 estimates, based on all possible samples of size n extracted from one population. Studying

123 the sampling distribution is very useful, as it allows to assess the goodness of an effect size

124 estimator and more specifically, three desirable properties of a good estimator for inferentiel

125 purposes: **unbiasedness**, **consistency** and **efficiency**.

126   An estimator is unbiased if the distribution of estimates is centered around the true

127 population parameter. On the other hand, an estimator is positively (or negatively) biased if

128 the distribution is centered around a value that is higher (or smaller) than the true

129 populatione parameter (see Figure 1). In other words, the bias tells us if estimates are good,

130   on average. The *bias* of a point estimator $\hat{\delta}$ can be computed as follows:

$$\hat{\delta}_{bias} = E(\hat{\delta}) - \delta \tag{1}$$

131   Where $E(\hat{\delta})$ is the mean expectency of the sampling distribution of the estimator and δ
132   is the true parameter.

133   In order to compare the *bias* of a point estimator for different true population
134   parameters, we can compute the bias divided by δ.

$$\hat{\delta}_{bias} = \frac{E(\hat{\delta}) - \delta}{\delta} \tag{2}$$

135   Bias informs us about the goodness of estimates averages, but says nothing about
136   individual estimates. Imagine a situation where the distribution of estimates is centered
137   around the real parameter but with such a large variance that some point estimates are very
138   far from the center. It would be problematic, as long as we have only one estimate, the one
139   based on our sample, and we don't know how far is this estimate from the center of the
140   sampling distribution. We hope that *all* possible estimates are close enough of the true
141   population parameter, in order to be sure that for *any* estimate, one has a correct estimation
142   of the real parameter. In other words, we expect the variability of estimates around the true
143   population parameter to be as small as possible. It refers to the **efficiency** of the point
144   estimator $(\hat{\delta})$ and can be computed as follows:

$$\hat{\delta}_{efficiency} = Var(\hat{\delta}) \tag{3}$$

145   Among all unbiased estimators, the more efficient will be the one with the smallest
146   variance.

147   Note that both unbiasedness and efficiency are very important. Remember that we

148 hope that *any* possible estimate is close of the real parameter. An unbiased estimator with

149 such a large variance that somes estimates are extremely far from the real parameter is as

150 undesirable as a parameter which is highly biased. In some situations, it is better to have a

151 very slightly biased estimator with a tigh shape around the biased value, so each estimate

152 "misses" the real parameter a little, than an unbiased estimator with a large variance [Ref to

153 add: https://eranraviv.com/bias-vs-consistency/]. Because both *unbiasedness* and *efficiency*

154 must be considered, it is interesting to compute an indicator that take simultaneously both

155 properties into account (Wackerly, Mendenhall, & Scheaffer, 2008). The *mean square error*

156 of a point estimator $\hat{\delta}$ is defined as follows:

$$MSE(\hat{\delta}) = E[(\hat{\delta} - \delta)^2] \tag{4}$$

157   It can be proven that the *mean square error* is a function of the bias and the variance

158 of $\hat{\delta}$:

$$MSE(\hat{\delta}) = \hat{\delta}_{efficiency} + \hat{\delta}^2_{bias} \tag{5}$$

159   Finally, the last property of a good point estimator is **consistency**: consistency means

160 that the bigger the sample size, the closer the estimate of the population parameter. In other

161 words, the estimates *converge* to the true population parameter.

Other purposes

163   According to Cumming (2013), an effect size estimator need to have a constant value

164 across designs in order to be easily interpretable and to be included in meta-analysis. In

165 other word, it should achieve the property of **generality**.

Different measures of effect sizes

166

167   The *d*-family effect sizes are commonly used with "between-subject" designs where

168 individuals are randomly assigned into one of two independant groups and groups scores

169 means are compared. In this setting, the inferential purpose of effect size measure is crucial.

170 We will therefore first focus our attention on the inferential purposes and then adress the

171 comparative and interpretative purposes as a second step.

172   The population effect size is defined as follows:

$$\delta = \frac{\mu_1 - \mu_2}{\delta} \tag{6}$$

173   They exist different estimators of this effect size measure varying as a function of the

174 chosen standardizer (δ). For all estimators, the mean difference is estimated by the difference

175 of both sample means ($\bar{X}_1 - \bar{X}_2$). When used for inference, some of them rely on both

176 assumptions of normality and equality of variances, while others rely solely on the normality

177 assumption.

Alternatives when variances are equal between groups

178

179   The most common estimator of δ is Cohen's d_s where the sample mean difference is

180 divided by a pooled error term (Cohen, 1965):

$$Cohen's d_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1) \times SD_1 + (n_2-1) \times SD_2}{n_1 + n_2 - 2}}} \tag{7}$$

181   The reasoning behind this measure is that considering both samples as extracted from

182 a common population variance (n.d.), we achieve a more accurate estimation of the

183 population variance by pooling both estimates of this parameter (i.e SD_1 and SD_2) and

184 because the larger the sample size, the more accurate the estimate, we give more weight to

¹⁸⁵ the estimate based on the larger sample size ($max(n_j)$). Unfortunately, even under the

¹⁸⁶ assumptions that residuals are independant and identically normally distributed with the

¹⁸⁷ same variance across groups, Cohen's d_s is known to be positively biased [Thompson, 2006,

¹⁸⁸ cited by Lakens, 2013] and for this reason, Hedges & Olkin (1985) has defined a

¹⁸⁹ bias-corrected version, which is referred to:

$$Hedge'sg_s = Cohen'sd_s \times (1 - \frac{3}{4 \times (n_1 + n_2) - 9}) \tag{8}$$

¹⁹⁰ The pooled SD is the best choice when variances are equal between groups (Grissom &

¹⁹¹ Kim, 2001) but they may not be well advised for use with data that violates this assumption

¹⁹² (Cumming, 2013; Grissom & Kim, 2001, 2005; Kelley, 2005, 2005; Shieh, 2013). In case of a

¹⁹³ positive pairing (i.e. the group with the larger sample size also has the larger variance), the

¹⁹⁴ variance will be over-estimated and therefore, the estimator will be lower as it should be. On

¹⁹⁵ the other side, in case of negative pairing (i.e. the group with the larger sample size has the

¹⁹⁶ smaller variance), the estimator will be larger as it should be. However, this assumption is

¹⁹⁷ very rare in practice (Cain, Zhang, & Yuan, 2017; Delacre, Lakens, & Leys, 2017; Delacre,

¹⁹⁸ Leys, Mora, & Lakens, 2019; Erceg-Hurn & Mirosevich, 2008; Glass, Peckham, & Sanders,

¹⁹⁹ 1972; Grissom, 2000; Micceri, 1989; Yuan, Bentler, & Chan, 2004). For this reason, both

²⁰⁰ Cohen's d_s and Hedge's g_s should be abandoned in favor of a robust alternative to unequal

²⁰¹ variances.

²⁰² **Alternatives when variances are unequal between groups**

²⁰³ In his review, Shieh (2013) mention three alternative available in the literature: the

²⁰⁴ sample mean difference, divided by the non pooled average of both variance estimates (A),

²⁰⁵ the Glass's d_s (B) and the Shieh's d_s (C).

²⁰⁶ The **sample mean difference, divided by the non pooled average of both**

²⁰⁷ **variance estimates** (A) was suggested by Cohen (1988). We immediately exclude this

208 alternative because it suffers of many limitations:

209   - it results in a variance term of an artificial population and is therefore very difficult

210 to interpret (Grissom & Kim, 2001);

211   - unless both sample sizes are equal, the variance term does not correspond to the

212 variance of the mean difference (Shieh, 2013);

213   - unless the mean difference is null, the measure is biased. Moreover, the bigger the

214 sample size, the larger the variance around the estimate.

215   **Glass's d_s.** When comparing one control group with one experimental group, Glass,

216 McGav, & Smith (2005) recommend using the SD of the control group as standardizer. It is

217 also advocated by (Cumming, 2013), because according to him, it is what makes the most

218 sense, conceptually speaking.

$$Glass'sd_s = \frac{\bar{X}_{experimental} - \bar{X}_{control}}{SD_{control}} \tag{9}$$

219   Because the SD of the experimental group has no impact on the computed Glass's d_s,

220 one could advice to report both mean differences divided by the SD of the control group and

221 mean differences divided by the SD of the experimental groups. However, it could induces

222 large ambiguity because both measures could be substantially different (Shieh, 2013).

223   **Shieh's d.** Kulinskaya and Staudte (2007) adviced the use of a standardizer that

224 take the sample sizes allocation ratios into account, in addition to the variance of both

225 groups. It results in a modification of the exact *SD* of the sample mean difference:

$$Shieh'sd_s = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{SD_1^2/q_1 + SD_2^2/q_2}} \tag{10}$$

226   We already mentioned that such a standardizer is required for inferential purposes, in

227 the context of heteroscedasticity. At the same time, the lack of generality caused by taking

228 this specificity of the design into account has led Cumming (2013) to question its usefulness

229 in terms of interpretability: when keeping constant the mean difference as well as SD_1 and

230 SD_2, Shieh's d_s will vary as a function of the sample sizes allocation ratio (dependency of

231 Shieh's d_s value on the sample sizes allocation ratio is detailed and illustrated in Appendix 1,

232 and also in the following shiny application: http: //127.0.0.1:3461/). This paradox would be

233 resolved if it were possible to answer the following question: "for any sample sizes allocation

234 ratio, is it possible to transform Shieh's d_s value in order to know what value of Shieh's d_s

235 would have been obtained if design were balanced (i.e. $n_1 = n_2$)?" Fortunately, we found

236 such a transformation. It can be shown that there is a relationship between the Shieh's d_s

237 value when samples sizes are equal between groups (i.e. $\delta_{Shieh,n1=n2}$) and Shieh's d_s values for

238 all other sample sizes allocation ratios

$$\delta_{Shieh,n_1=n_2} = \delta_{Shieh} \times \frac{(nratio + 1) \times \sigma_{n_1 \neq n_2}}{2 \times \sigma_{n_1=n_2} \times \sqrt{nratio}} \tag{11}$$

239   With

$$\sigma_{n_1=n_2} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}$$

240 and

$$\sigma_{n_1 \neq n_2} = \sqrt{(1 - \frac{n_1}{N}) \times \sigma_1^2 + (1 - \frac{n_2}{N}) \times \sigma_2^2}$$

241   Thanks to this equation, Shieh's δ can be compared across two different studies using

242 different sample sizes allocation ratio and could be included in meta-analysis.

243 **Simulations**

244   We performed Monte Carlo simulations using R (version 3.5.0) to assess the bias,

245 efficiency and consistency of the 6 measures of effect sizes described in previous section:

246 Cohen's d_s, Hedge's g_s, Glass's d_s using respectively the sample standard deviation of the

247 first or second group as a standardizer, Shieh's d_s and our transformed measure of Shieh's d_s,

248 that we will note later d_s^*.

₂₄₉   100,000 datasets were generated for 1,260 scenarios. In 315 scenarios, samples were

₂₅₀ extracted from a normal population distribution and in 945 scenarios, samples were

₂₅₁ extracted from non normal population distributions. In order to assess the goodness of

₂₅₂ estimators under realistic deviations from the normality assumption, we referred to the

₂₅₃ review of Cain et al. (2017). Based on their investigation[5], Cain et al. (2017) found values of

₂₅₄ kurtosis from G2 = -2.20 to 1,093.48. According to their suggestions, throughout our

₂₅₅ simulations, we kept constant the population kurtosis value at the 99th percentile of their

₂₅₆ distribution, i.e. G2=95.75. Regarding skewness, we simulated population parameter values

₂₅₇ which correspond to the 1st and 99th percentile of their distribution, i.e. respectively G1 =

₂₅₈ -2.08 and G1 = 6.32. We also simulated null population parameter values (i.e. G1 = 0), in

₂₅₉ order to assess the main effect of high kurtosis on the goodness of estimators. All possible

₂₆₀ combinations of skewness and kurtosis and the number of scenarios for each combination are

₂₆₁ summarized in Table 1.

₂₆₂   Table 1. *Number of Combinations of skewness and kurtosis in our simulations*

₂₆₃   *Note.* Fisher's skewness (G1) and kurtosis (G2) are presented in Table 1. The 315

₂₆₄ combinations where both G1 and G2 equal 0 correspond to the normal case.

₂₆₅   For the 4 resulting combinations of skewness and kurtosis (see Table 1), all other

₂₆₆ population parameter values were chosen in order to illustrate the consequences of factors

₂₆₇ known to play a key role on goodness of estimators. We manipulated the sample sizes, the

₂₆₈ sample size ratio (n-ratio = $\frac{n_1}{n_2}$), the *SD*-ratio (*SD*-ratio = $\frac{\sigma_1}{\sigma_2}$), and the sample size and

₂₆₉ variance pairing. In our scenarios, μ_2 was always 0 and μ_1 varied from 0 to 4, in step of 1 (so

₂₇₀ does $\mu_1 - \mu_2$). Moreover, σ_1 always equals 1, and σ_2 equals .1, .25, .5, 1, 2, 4 or 10 (so does

---

[5] Cain et al. (2017) investigated 1,567 univariate distributions from 194 studies published by authors in

Psychological Science (from January 2013 to June 2014) and the *American Education Research Journal*

(from January 2010 to June 2014). For each distribution, they computed the Fisher's skewness (G1) and

kurtosis (G2).

the *SD*-ratio). The simulations for which both σ_1 and σ_2 equal 1 are the particular case of homoscedasticity (i.e. equal variances across groups). Sample size of both groups (n_1 and n_2) were 20, 50 or 100. When sample sizes of both groups are equal, the *n*-ratio equals 1 (it is known as a balanced design). All possible combinations of *n*-ratio and *SD*-ratio were performed in order to distinguish positive pairings (the group with the largest sample size is extracted from the population with the largest *SD*), negative pairings (the group with the smallest sample size is extracted from the population with the smallest *SD*), and no pairing (sample sizes and/or population *SD* are equal across all groups). In sum, the simulations grouped over different sample sizes yield 5 conditions based on the *n*-ratio, *SD*-ratio, and sample size and variance pairing, as summarized in Table 2.

Table 2. *5 conditions based on the n-ratio, SD-ratio, and sample size and variance pairing*

Note. The *n*-ratio is the sample size of the first group (n_1) divided by the sample size of the second group (n_2). When all sample sizes are equal across groups, the *n*-ratio equals 1. When $n_1 > n_2$, *n*-ratio > 1, and when $n_1 < n_2$, *n*-ratio < 1. *SD*-ratio is the population *SD* of the first group (σ_1) divided by the population *SD* of the second group (σ_2). When $\sigma_1 = \sigma_2$, *SD*-ratio $= 1$. When $\sigma_1 > \sigma_2$, *SD*-ratio > 1. Finally, when $\sigma_1 < \sigma_2$, *SD*-ratio < 1.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in aera publications. *Educational Researcher*, *35*, 33–40. doi:10.3102/0013189X035006033

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

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology, 30*(1), 92–101. doi:10.5334/irsp.82

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of welch's f-test instead of the classical f-test in one-way anova. *International Review of Social Psychology, 32*(1), 1–12. doi:http://doi.org/10.5334/irsp.198

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

319   Glass, G. V., McGav, B., & Smith, M. L. (2005). *Meta-analysis in social research*
320 (Sage.). Beverly Hills, CA.

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

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

326   Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the
327 appropriate conceptualization of effect size. *Psychological Methods, 6*(2), 135–146.
328 doi:10.1037/1082-989X.6.2.135

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

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

333   Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis* (Academic
334 Press.). Cambridge, Massachusetts. doi:10.1016/C2009-0-03396-0

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

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

341   Kirk, R. E. (2009). Practical significance: A concept whose time has come. *Educational*

342 *and Psychological Measurement, 56*(5), 746–759. doi:10.1177/0013164496056005002

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

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

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

351   Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian
352 defense and two principles that warrant it. *Psychological Inquiry, 1*(2), 108–141.

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

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

358   Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies:
359 Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*,
360 241–286. doi:10.1006/ceps.2000.1040

361   Peng, C.-Y., & Chen, L.-T. (2014). Beyond cohen's d: Alternative effect size measures
362 for between-subject designs. *THE JOURNAL OF EXPERIMENTAL EDUCATION, 82*(1),
363 22–50. doi:10.1080/00220973.2012.745471

364   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

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics
with applications (7th edition)* (Brooks/Cole, Cengage Learning.). Belmont, USA.

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

(n.d.).

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

empirical social science research in aera publications. *Educational Researcher*, *35*, 33–40. doi:10.3102/0013189X035006033

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

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use welch's t-test instead of student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. doi:10.5334/irsp.82

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of welch's f-test instead of the classical f-test

410 in one-way anova. *International Review of Social Psychology*, *32*(1), 1–12.

411 doi:http://doi.org/10.5334/irsp.198

412   Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods:

413 An easy way to maximize the accuracy and power of your research. *American Psychologist*,

414 *63*(7), 591–601. doi:10.1037/0003-066X.63.7.591

415   Fan, X. (2001). Statistical significance and effect size in education research: Two sides

416 of a coin. *Journal of Educational Research*, *94*(5), 275–282. doi:10.1080/00220670109598763

417   Glass, G. V., McGav, B., & Smith, M. L. (2005). *Meta-analysis in social research*

418 (Sage.). Beverly Hills, CA.

419   Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet

420 assumptions underlying the fixed effects analyses of variance and covariance. *Review of*

421 *Educational Research*, *42*(3), 237–288. doi:10.3102/00346543042003237

422   Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting*

423 *and Clinical Psychology*, *68*(1), 155–165. doi:10.1037//0022-006x.68.1.155

424   Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the

425 appropriate conceptualization of effect size. *Psychological Methods*, *6*(2), 135–146.

426 doi:10.1037/1082-989X.6.2.135

427   Grissom, R. R., & Kim, J. J. (2005). *Effect size for research: A broad practical*

428 *approach.* (Lawrence Erlbaum Associates, Mahwah, N.J.). London.

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

430 York.

431   Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis* (Academic

432 Press.). Cambridge, Massachusetts. doi:10.1016/C2009-0-03396-0

433 Henson, R. I., & Smith, A. D. (2000). State of the art in statistical significance and

434 effect size reporting: A review of the APA task force report and current trends. *Journal of*

435 *Research and Development in Education, 33*(4), 285–296.

436 Kelley, K. (2005). The effects of nonnormal distributions on confidence intervales

437 around the standardized mean difference: Bootstrap and parametric confidence intervals.

438 *Educational and Psychological Measurement, 65*(1), 51–69. doi:10.1177/0013164404264850

439 Kirk, R. E. (2009). Practical significance: A concept whose time has come. *Educational*

440 *and Psychological Measurement, 56*(5), 746–759. doi:10.1177/0013164496056005002

441 Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative

442 science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*(863), 1–12.

443 doi:10.3389/fpsyg.2013.00863

444 Li, J. (2016). Effect size measures in a two-independent-samples case with nonnormal

445 and nonhomogeneous data. *Behavior Research Methods, 48*(4), 1560–1574.

446 doi:10.3758/s13428-015-0667-z

447 McBride, G. B., Loftis, J. C., & Adkins, N. C. (1993). What do significance tests really

448 tell us about the environment? *Environmental Management, 17*(4), 423–432.

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

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

451 Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures.

452 *Psychological Bulletin, 105*(1), 156–166. doi:10.1037/0033-2909.105.1.156

453 Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical

454 significance: A practical guide for biologists. *Biological Reviews, 82*, 591–605.

455 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

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics with applications (7th edition)* (Brooks/Cole, Cengage Learning.). Belmont, USA.

479 Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in

480 psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

481 Yuan, K.-H., Bentler, P. M., & Chan, W. (2004). Structural equation modeling with

482 heavy tailed distributions. *Psychometrika*, *69*(3), 421–436. doi:10.1007/bf02295644

483 (n.d.).

		Kurtosis		
		0	95.75	**TOTAL**
	0	315	315	**630**
Skewness	-2.08	/	315	**315**
	6.32	/	315	**315**
	TOTAL	**315**	**945**	**1260**

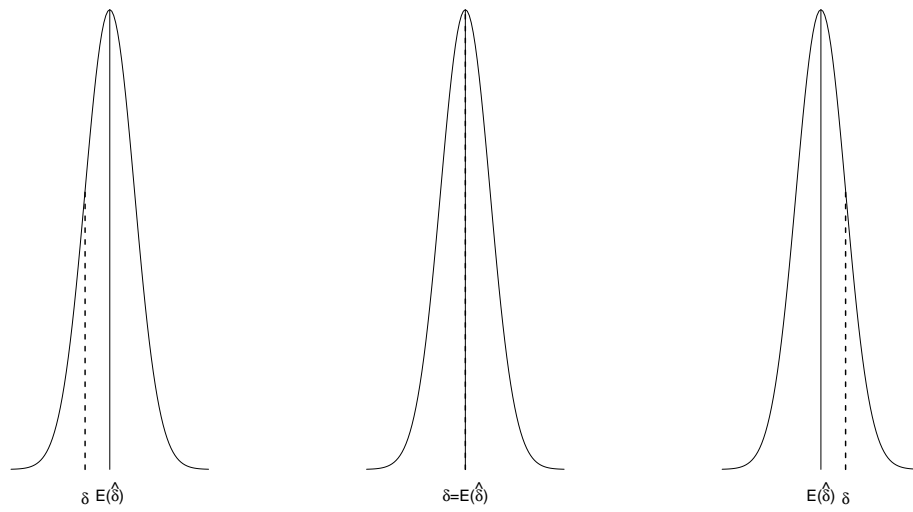|		\(n\)-ratio		
		1	**>1**	**<1**
	1	a	b	b
SD-ratio	**>1**	c	d	e
	<1	c	e	d

Figure 1. Samplig distribution for a positively biased (left), an unbiased (center) and a negatively biased estimator (right)

Appendix

Appendix 1: The mathematical study of Shieh's δ

Paste Appendix 1 when it will be finished

Appendix 2: Confidence intervals

Paste Appendix 2 when it will be finished

Appendix 3: a priori power analyses

Paste Appendix 3 when it will be finished (Cumming & Finch, 2001)

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervales that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*(532), 532–574.