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

7 Abstract

8

9 *Keywords:* keywords

10 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 [TROUVER UN TERME] significance (e.g. clinical, personnal, social, professionnal)

³⁵ of a result (i.e. the relevance of an effect in real life). Moreover, there are several situations

³⁶ that call for effect size measures and in the current litterature, it's not always easy to know

³⁷ which measure using in a specific context. The first aim of this paper is therefore twofold:

³⁸   1. Clearly define what is (and what is not) a measure of effect size;

³⁹   2. Listing the different situations that call for effect sizes measure and define required

⁴⁰ properties as a function of the situations;

⁴¹   Second, many differents estimators of effect sizes are available in literature and it is not

⁴² always easy to know which measures is appropriate in which circumstance. We will limit our

⁴³ study to "between-subject" designs where individuals are randomly assigned into one of two

⁴⁴ independant groups and groups scores are compared based on their means. More specifically,

⁴⁵ we will focus on the standardized mean difference, called the d-family, because it is the

⁴⁶ dominant family of estimators of effect size when comparing two groups based on their means

⁴⁷ (Peng, Chen, Chiang, & Chiang, 2013; Shieh, 2013). We will see that even in this very

⁴⁸ specific context, there is little agreement between researchers as to which is the most suitable

⁴⁹ estimator. According to us, the main reason is that it is difficult to find a measure which

⁵⁰ optimally serves all the purposes of an effect size measure. For example, interpretability is

⁵¹ obtained at the expense of inferential properties and vice versa. The second aim of this paper

⁵² is therefore to review the most famous estimators of this family next to the role they serve.

⁵³   **Measure of effect size: what it is, what it is not**

⁵⁴   The effect size is commonly refered to the practical significance of a test. Grissom and

⁵⁵ Kim (2005) define the effect size as the extent to which results differ from what is implied by

⁵⁶ the null hypothesis. In the context of the comparison of two groups based on their mean,

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

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

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

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

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

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

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

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

65   First, benchmarks about what should be a small, medium or large effect size might

66 have contributed at seeing the effect size as a measure of the importance or the relevance of

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

68 mathematical indicator of the magnitude of a difference, which depends on the way a

69 variable is converted into numerical indicator. In order to assess the meaningfulness of an

70 effect, we should be able to relate this effect with behaviors/meaningful consequences in the

71 real world (Andersen, McCullagh, & Wilson, 2007). For example, let us imagine a sample of

72 students in serious school failure who are randomly divided into two groups: an experimental

73 group following a training program and a control group. At the end of the training, students

74 in the experimental group have on average significantly higher scores on a test than students

75 in the control group, and the difference is large (e.g. 30 percents). Does it mean that

76 students in the experimental condition will be able to pass to the next grade and to continue

77 normal schooling? Whether the computed magnitude of difference is an important,

78 meaningful change in everyday life refers to another construct: the *??? significance* (Bothe

79 & Richardson, 2011). It refers to the interpretation of treatment outcomes and is neither

80 statistical nor mathematical, it is related to underlying theory that posits an empirical

81 hypothesis. In other words, the relation between *practical* and *???* significance is more a

82 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)

When used for inference, effect size measures are generally submitted to a range of assumptions (e.g. independent and identically distributed residuals are normal and have equal variances between groups). When these assumptions are not met, many estimations of effect size are inaccurate and alter the robustness of the statistical conclusions.

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:

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

- To incorporate results in meta-analysis (Lakens, 2013; Li, 2016; Nakagawa & Cuthill, 2007; Stout & Ruble, 1995; Wilkinson & the Task Force on Statistical Inference, 1999)

Finally, effect size measures can be used for *interpretive* purposes: in order to assess the practical significance of a result (beyond statistical significance; Lakens, 2013; American

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

109   **Robust measures**

110 **Properties of a good effect size estimator (for inferential purposes)**

111   The value of the estimate of an estimator depends on the sampling. That is to say,

112 based on different samples extracted from the same population, one would obtain different

113 estimates of the same estimator. The *sampling distribution* of the estimator is the

114 distribution of all estimates, based on all possible samples of size n extracted from one

115 population. Studying the sampling distribution is very useful, as it allows to assess the

116 goodness of an effect size estimator and more specifically, three desirable properties of a good

117 estimator: **unbiasedness**, **consistency** and **efficiency**.

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

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

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

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

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

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

123   Where $E(\hat{\delta})$ is the mean of the sampling distribution of the estimator and δ is the true

124 parameter.

125   In order to compare the *bias* of a point estimator for different true population

126 parameters, we can compute the bias divided by δ.

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

¹²⁷ Bias informs us about the goodness of estimates averages, but says nothing about

¹²⁸ individual estimates. Imagine a situation where the distribution of estimates is centered

¹²⁹ around the real parameter but with such a large variance that some point estimates are very

¹³⁰ far from the center. It would be problematic, as long as we have only one estimate, the one

¹³¹ based on our sample, and we don't know how far is this estimate from the center of the

¹³² sampling distribution. We hope that *all* possible estimates are close enough of the true

¹³³ population parameter, in order to be sure that for *any* estimate, one has a correct estimation

¹³⁴ of the real parameter. In other words, we expect the variability of estimates around the true

¹³⁵ population parameter to be as small as possible. It refers to the **efficiency** of the point

¹³⁶ estimator ($\hat{\delta}$) and can be computed as follows:

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

¹³⁷ Among all unbiased estimators, the more efficient will be the one with the smallest

¹³⁸ variance.

¹³⁹ Note that both unbiasedness and efficiency are very important. Remember that we

¹⁴⁰ hope that *any* possible estimate is close of the real parameter. An unbiased estimator with

¹⁴¹ such a large variance that somes estimates are extremely far from the real parameter is as

¹⁴² undesirable as a parameter which is highly biased. In some situations, it is better to have a

¹⁴³ very slightly biased estimator with a tigh shape around the biased value, so each estimate

¹⁴⁴ "misses" the real parameter a little, than a biased estimator with a large variance [Ref to

¹⁴⁵ add: https://eranraviv.com/bias-vs-consistency/]. Because both *unbiasedness* and *efficiency*

¹⁴⁶ must be considered, it is interesting to compute an indicator that take simultaneously both

¹⁴⁷ properties into account (Wackerly, Mendenhall, & Scheaffer, 2008). The *mean square error*

¹⁴⁸ of a point estimator $\hat{\delta}$ is defined as follows:

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

149 It can be proven that the *mean square error* is a function of the bias and the variance
150 of $\hat{\delta}$:

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

151 Finally, the last property of a good point estimator is **consistency**: consistency means
152 that the bigger the sample size, the closer the estimate of the population parameter. In other
153 words, the estimates *converge* to the true population parameter.

154 **Properties of a good effect size estimator (for comparative and interpretive**
155 **purposes)**

156 **Generality**

157 According to Cumming (2013), an effect size estimator need to have a constant value
158 across designs in order to be easily interpretable and to be included in meta-analysis.

159 At first glance, this quality is incompatible with the mathematical properties required
160 for an inferential purpose. For example, according to the statistical properties of Welch's
161 statistic, in the context of heteroscedasticity, it seems required to take the sample size
162 allocation ratio into account in order to define a proper inferential measure of effect size
163 (**???**). It would mean that for the same amount of differences between two means, same
164 standard deviations and σ-ratio, a proper effect size estimator would vary as a function of
165 the sample sizes allocation ratio, which would make it very dependent on the characteristics
166 of the design and therefore very difficult to interpret.

167 In the following section, we will review the most popular effect size measures in the *d*

168 family. In this section, we will also propose an original transformation of the Shieh's δ. This

169 transformation should help at bridging the imperatives of generality and need to take the

170 specificities of the design into account.

Different measures of effect sizes

172   As previously mentioned, we will limit our study to the d-family, commonly used with

173 "between-subject" designs where individuals are randomly assigned into one of two

174 independant groups and groups scores are compared based on their means. We will first

175 focus our attention on the inferential purpose of the effect size measures.

176   The population effect size is defined as follows:

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

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

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

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

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

181 assumption.

Alternatives when variances are equal between groups

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

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

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

185   The reasoning behind this measure is that considering both sample as extracted from a

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

187 variance by pooling both estimates of this parameter (i.e SD_1 and SD_2) and because the

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

189 based on the larger sample size ($max(n_j)$). Unfortunately, even under the assumptions that

190 residuals are independant and identically normally distributed with the same variance across

191 groups, Cohen's d_s is known to be positively biased [Thompson, 2006, cited by Lakens, 2013]

192 and for this reason, Hedges and Olkin (1985) has defined a bias-corrected version, which is

193 referred to:

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

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

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

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

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

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

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

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

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

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

203 1972; Grissom, 2000; Micceri, 1989; Yuan, Bentler, & Chan, 2004). For this reason, both

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

205 variances.

Alternatives when variances are unequal between groups

207   While it is becoming more common in statistical software to present Welch's t-test by

208 default, when performing a t-test (i.e., R, Minitab), similar issues for the measures of effect

209 sizes has received less attention (Shieh, 2013) and Cohen's d_s remains persistent [2]. One

210 possible reason is that researchers cannot find a consensus on which alternative should be in

211 use (Shieh, 2013). In his review, Shieh (2013) mention three alternative available in the

212 literature: the sample mean difference, divided by the non pooled average of both variance

213 estimates (A), the Glass's d_s (B) and the Shieh's d_s (C).

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

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

216 alternative because it suffers of many limitations:

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

218 to interpret (Grissom & Kim, 2001); - unless both sample sizes are equal, the variance term

219 does not correspond to the variance of the mean difference (Shieh, 2013);

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

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

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

223 McGav, and Smith (2005) recommend using the SD of the control group as standardizer. It

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

225 sense, conceptually speaking.

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

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

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

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

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

————

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

Shieh's d.   Kulinskaya and Staudte (2007) adviced the use of a standardizer that take the sample sizes allocation ratios into account, in addition to the variance of both groups. It results in a modification of the exact SD of the sample mean difference:

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

We already mentioned that such a standardizer is required for inferential purposes, in the context of heteroscedasticity. At the same time, the lack of generality caused by taking this specificity of the design into account has led Cumming (2013) to question its usefulness in terms of interpretability: when keeping constant the mean difference as well as SD_1 and SD_2, Shieh's d_s will vary as a function of the sample sizes allocation ratio (dependency of Shieh's d_s value on the sample sizes allocation ratio is detailed and illustrated in Appendix 1, and also in the following shiny application: http: //127.0.0.1:3461/). This paradox would be resolved if it were possible to answer the following question: "for any sample sizes allocation ratio, is it possible to transform Shieh's d_s value in order to know what value of Shieh's d_s would have been obtained if design were balanced (i.e. $n_1 = n_2$)?" Fortunately, such a transformation is possible. It can be shown that there is a relationship between the Shieh's d_s value when samples sizes are equal between groups (i.e. $\delta_{Shieh,n1=n2}$) and Shieh's d_s values for all other sample sizes allocation ratios

$$\delta_{Shieh,n_1=n_2} = \delta_{Shieh} \times \frac{\frac{\mu_1-\mu_2}{2\times\sigma_{(n_1=n_2)}}}{\frac{(\mu_1-\mu_2)\times\sqrt{nratio}}{(nratio+1)\times\sigma_{(n_1\neq n_2)}}} \leftrightarrow \delta_{Shieh,n_1=n_2} = \delta_{Shieh} \times \frac{(nratio+1)\times\sigma_{n_1\neq n_2}}{2\times\sigma_{n_1=n_2}\times\sqrt{nratio}} \tag{11}$$

With

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

and

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

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

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

Simulations

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

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

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

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

255 that we will note later d_s^*.

256   100 000 datasets were generated for 1260 scenarios. In 315 scenarios, samples were

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

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

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

260 review of Cain et al. (2017). Based on their investigation [3], Cain et al. (2017) found values

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

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

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

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

265 -2.08 and G1 = 6.32. We also simulated null population parameter values (i.e. G1 = 0), in

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

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

268 summarized in Table 1.

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

---

[3] 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).

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

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

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

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

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

275   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

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

277   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

278   the SD-ratio). The simulations for which both σ_1 and σ_2 equal 1 are the particular case of

279   homoscedasticity (i.e. equal variances across groups). Sample size of both groups (n_1 and n_2)

280   were 20, 50 or 100. When sample sizes of both groups are equal, the n-ratio equals 1 (it is

281   known as a balanced design). All possible combinations of n-ratio and SD-ratio were

282   performed in order to distinguish positive pairings (the group with the largest sample size is

283   extracted from the population with the largest SD), negative pairings (the group with the

284   smallest sample size is extracted from the population with the smallest SD), and no pairing

285   (sample sizes and/or population SD are equal across all groups).

286   In sum, the simulations grouped over different sample sizes yield 5 conditions based on

287   the n-ratio, SD-ratio, and sample size and variance pairing, as summarized in Table 2.

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

289   *pairing*

290   *Note.* The n-ratio is the sample size of the first group (n_1) divided by the sample size

291   of the second group (n_2). When all sample sizes are equal across groups, the n-ratio equals 1.

292   When $n_1 > n_2$, n-ratio > 1, and when $n_1 < n_2$, n-ratio < 1. SD-ratio is the population SD

293   of the first group (σ_1) divided by the population SD of the second group (σ_2). When

294   $\sigma_1 = \sigma_2$, *SD*-ratio = 1. When $\sigma_1 > \sigma_2$, *SD*-ratio > 1. Finally, when $\sigma_1 < \sigma_2$, *SD*-ratio < 1.

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

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

297   doi:10.3102/0013189X035006033

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

299   *psychological association [apa] (6 ed.)* (American Psychological Association.). Washington,

300   DC:

301   Andersen, M. B., McCullagh, P., & Wilson, G. J. (2007). But what do the numbers

302   really tell us? Arbitrary metrics and effect size reporting in sport psychology research.

303   *Journal of Sport & Exercise Psychology, 29*, 664–672.

304   Bothe, A. K., & Richardson, J. D. (2011). Statistical, practical, clinical, and personal

305   significance: Definitions and applications in speech-language pathology. *American Journal of*

306   *Speech-Language Pathology, 20*, 233–242.

307   Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness

308   and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior*

309   *Research Methods, 49*(5), 1716–1735. doi:10.3758/s13428-016-0814-1

310   Cohen, J. (1965). Some statistical issues in psychological research. In *Handbook of*

311   *clinical psychology* (B. B. Wolman., pp. 95–121). New York: McGraw-Hill.

312   Cumming, G. (2013). Cohen's d needs to be readily interpretable: Comment on shieh

313   (2013). *Behavior Research Methods, 45*, 968–971. doi:10.3758/s13428-013-0392-4

314   Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use

315   welch's t-test instead of student's t-test. *International Review of Social Psychology, 30*(1),

316   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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

352 doi:10.3389/fpsyg.2013.00863

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

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

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

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

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

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

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

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

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

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

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

364 doi:10.1111/j.1469-185X.2007.00027.x

365   Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies:

366 Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*,

367 241–286. doi:10.1006/ceps.2000.1040

368   Peng, C.-Y., & Chen, L.-T. (2014). Beyond cohen's d: Alternative effect size measures

369 for between-subject designs. *THE JOURNAL OF EXPERIMENTAL EDUCATION*, *82*(1),

370 22–50. doi:10.1080/00220973.2012.745471

371   Peng, C.-Y., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The Impact of APA

372 and AERA Guidelines on Effect size Reporting. *Contemporary Educational Psychology*,

373 *82*(1), 22–50. doi:10.1080/00220973.2012.745471

374   Prentice, D., & Miller, D. T. (1990). When small effects are impressive. *Psychological*

375 *Bulletin*, *112*(1), 160–164.

376   Shieh, G. (2013). Confidence intervals and sample size calculations for the weighted

377 eta-squared effect sizes in one-way heteroscedastic ANOVA. *Behavior Research Methods*,

378 *45*(1), 2–37. doi:10.3758/s13428-012-0228-7

379   Steyn, H. S. (2000). Practical significance of the difference in means. *Journal of*

380 *Industrial Psychology*, *26*(3), 1–3.

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

382 results in accounting education research: The use of effect size information. *Journal of*

383 *Accounting Education*, *13*(3), 281–298.

384   Sullivan, G., & Feinn, R. (2012). Using effect size—or why the p value is not enough.

385 *Journal of Graduate Medical Education*, 279–282. doi:10.4300/JGME-D-12-00156.1

386 Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2008). *Mathematical statistics*

387 *with applications (7th edition)* (Brooks/Cole, Cengage Learning.). Belmont, USA.

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

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

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

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

392 (n.d.).

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

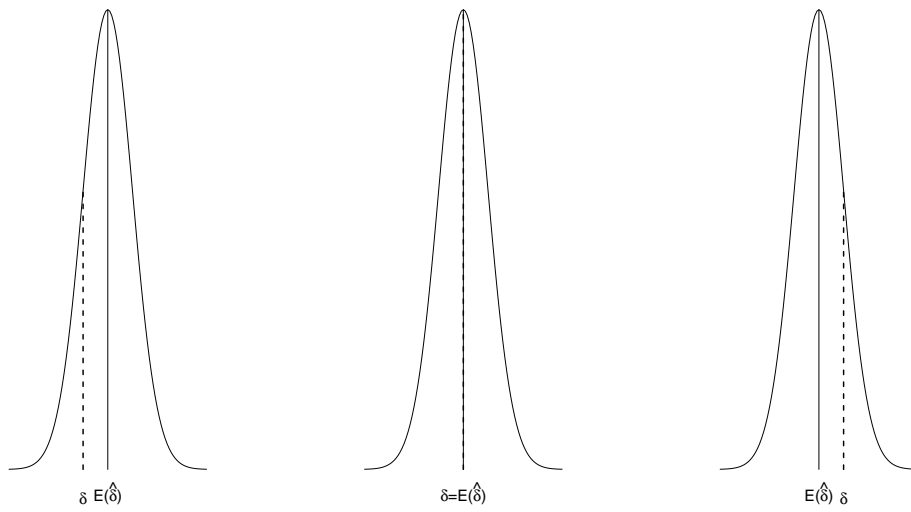|		\(n\)-ratio		
		1	**>1**	**<1**
	1	a	b	b
***SD*-ratio**	**>1**	c	d	e
	<1	c	e	d

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