Standardized means difference effect size measures for planned comparisons, trend analysis
and other applications of contrast analysis

Marcello Gallucci[1] & Marco Perugini[1]

[1] Department of Psychology, University of Milano-Bicocca

Abstract

Measures of effect size are increasingly important in planning, interpreting and reporting research results. In contrast analysis and its applications, such as planned comparisons and trend analysis, popular effect size indices are in the correlation metric, which may result difficult to interpret and to generalize across different designs. Here we discuss effect size indices based on standardized mean differences, inspired by the Cohen's d effect size measure. In contrast analysis, standardized mean differences can be interpreted and compared when they are scaled. Two scaling methods are discussed. The first method normalizes the contrast weights and makes the contrast effect size easy to use particularly in power analysis, because its value is strictly related with the inferential tests commonly employed in contrast analysis. A second method of scaling is proposed, which guarantees that the contrast effect size retains the same scale as the Cohen's d, across different designs and applications. Properties of the effect size measures, along with comparisons among different effect size measures are discussed. Practical advices and examples of computations are reported as well.

*Keywords:* Effect size, contrast analysis, power analysis

Standardized means difference effect size measures for planned comparisons, trend analysis

and other applications of contrast analysis

Measures of effect size have become an important tool for research in psychology, both for reporting and disseminating empirical results, and for planning new research (Cumming, 2014). Empirical effects can be expressed in terms of standardized effect size indices, with the advantage that they can be interpreted and compared independently of the variables measurement scale, often across different research designs and applications (Cohen, 1988). Despite the increasing attention given to effect size indices and their growing use (Kelley & Preacher, 2012), there are designs and analyses for which little progresses have been made. This is the case for effect size indices for contrast analysis, which is mostly based on a few seminal works (Cohen, 1988; Rosenthal, Rosnow, & Rubin, 2000) and occasional subsequent input (Furr, 2004; Liu, 2014; Steiger, 2004; Wahlsten, 1991). Contrast analysis is a statistical procedure characterized by focused tests of mean groups differences instead of omnibus tests of difference (Rosenthal & Rosnow, 1985). It can be understood as an instance of a model comparison perspective, insofar a specific contrast reflects a theoretical model (Judd, McClelland, & Ryan, 2011; Maxwell, Delaney, & Kelley, 2017).

Several authors have explained the benefits of this approach to data analysis, the most prominent of which is that it can provide direct evaluations of theoretically-driven predictions and hypotheses (Furr & Rosenthal, 2003). We wish to stress another important benefit of contrast analysis, namely that it requires focusing on the specific effect of interest which is a main pre-requiste for appropriate calculations of power analysis. Power analysis, and similar methods of a priori sample size planning, has gained considerable attention during the last few years. One important reason is the difficulty to replicate some results in Psychology, which has been explained also as a consequence of the combination of underpowered original studies and publication bias in the literature (Asendorpf et al., 2013; Bakker, Dijk, & Wicherts, 2012; Maxwell, 2004). One (positive) reaction to this state of affairs has been an increased emphasis on adequately powering one's study before starting

52  data collection. However, reaserchers willing to adopt a contrast analysis approach will find

53  difficulties in identifying a clear way to estimate contrast effect sizes from the literature.

54  Developed effect sizes in this area are expressed in a correlation metric (Rosenthal et al.,

55  2000). From correlation-based effect size is easy to convert to mean-based effect size indices,

56  yet their properties are not necessarily transferable and each has advantages and

57  disadvantages (McGrath & Meyer, 2006). Moreover, consider that contrast analysis involves

58  focused tests of groups means, hence the intuitive effect size unit is one that should involve

59  differences in means in its calculation (Wiens & Nilsson, 2017), such as Cohen's d and

60  related variants. It seems almost paradoxical then that most developments about effect size

61  indices in contrast analysis have been done in a correlational approach.

62      In this article we review measures of effect sizes for contrast analysis inspired by the

63  classical Cohens'd class of measures, with particular focus on their interpretability,

64  comparability, and inferential testability. By *interpretability* we mean the degree by which a

65  measure has a clear meaning, and thus a clear interpretation, when defined for different

66  statistical effects and applications. The increased attention for effect size measures in

67  Psychology has included a recommendation to report and interpret the magnitude of effects,

68  both in absolute and in relative terms (Henson & Smith, 2000; Wilkinson, 1999).

69  Interpreting an effect size index in absolute terms is much easier when it has a clear

70  definition and an intuitive metric. The ease of interpretation of an effect size index relatively

71  to the published results in a research field depends also on its degree of comparability. By

72  *comparability* (Bakeman, 2005; Glass, Smith, & McGaw, 1981; Keppel, 1991; Morris &

73  DeShon, 2002) we mean the degree by which a measure of effect size conveys the same

74  quantity when associated with the same effect, across different designs and applications.

75  Comparability is crucial expecially in power analysis, in which the researcher needs to guess

76  the expected effect size in order to compute power parameters (Beta, expected N, etc.).

77  When the expected effect size is inspired by published research, it is often the case that the

78  planned research is not exactly equal to the published one, thus the researcher needs to port,

79 i.e. translate, the observed effect size to the new design. A comparable and therefore portable

80 effect size makes this operation easier. Comparability is important also in meta-analysis:

81 When several effect size indices are gathered from the literature and are aggregated to

82 estimate in a reliable way the population effect sizes, the indices that are aggregated should

83 be comparable in the metrics and in the interpretation. If this is not the case, the overall

84 estimated effect would represent a mis-specification of the population parameters.

85 By *inferential testability* we mean the degree by which a measure of effect size is

86 directly associated with an inferential test used to test (null) hypotheses about the observed

87 results or to compute power in the planning phase of the research. There are effect size

88 indices, in fact, that can be readily associated with a statistical test, such has Cohen's d-like

89 measures (Cohen, 1988) and correlation indices (Rosenthal et al., 2000). This makes their

90 usage in power analysis and meta-analysis greatly facilitated. Other effect size measures are

91 not logically associated with an inferential test, thus their applicability may be limited to

92 specific phases of the research development. Inferential testability can be also evaluated for a

93 set of effect size measures. It is often the case, in fact, that an effect can be quantified with

94 different, alternative measures of size. Such a set of measures may, with varying degrees of

95 difficulty, be reconciled with the same inferential test, showing shared inferential testability.

96 This property allows to conduct coherent hypothesis testing and power analysis, and greatly

97 facilitates aggregation of results in meta-analyses. To provide an example, in regression

98 analysis an effect size can be quantified with several different measures, such as the

99 unstandardized $B$ coefficient, standardized $\beta$ coefficient, the $\eta^2$ or the partial $\eta^2$. All these

100 effect size measures express the same effect using different scales and emphasizing different

101 charateristics of the effect. However, they all share the same inferential test, usually the

102 t-test, and the same power function. In this article we present alternative definitions of effect

103 size measures for contrast analysis, with particular emphasis on their shared inferential

104 testability.

105 Finally, we also discuss practical aspects of computation and estimation of effect size

106 indices in contrast analysis, by discussing some theoretical results and presenting dedicated

107 software that can be used alongside popular software, such as R (R Core Team, 2008), SPSS

108 (IBM Corp., 2017), and G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). We accompany

109 the article with a R package named `cpower`, specifically developed to execute the statistical

110 procedures presented here. The R package `cpower` can be found at github[1].

## Contrasts analysis background

112   A contrast is a linear combination of means whose coefficients sum up to zero, meant

113 to estimate a particular comparison of means and test it against zero. We refer to the

114 contrast set of coefficients as $\boldsymbol{c} = \{c_i\}$, and to the expected set of means as $\boldsymbol{\mu} = \{\mu_i\}$. The

115 contrast coefficients (weights) are chosen such that $\sum_i c_i = 0$, with $i = \{1, .., k\}$ where $k$ is

116 the number of means being compared. The contrast expected value is $c\mu = \sum_i (c_i \cdot \mu_i)$. As

117 an example, consider a simple design with two groups: The comparison of the two groups

118 means can be carried out with a simple contrast with $\boldsymbol{c} = \{1, -1\}$, in which the contrast

119 value is simply the expected difference between means, $c\mu = c_1\mu_1 + c_2\mu_2 = \mu_1 - \mu_2$.

120   A contrast defined across $k$ means of independent groups of size $n$ can be tested

121 employing either an independent samples t-test or an F-test. The t-test expected value, with

122 $k(n-1)$ degrees of freedom, is (Steiger, 2004):

$$E(t_{k(n-1)}) = \frac{\sum (c_i \cdot \mu_i)}{\sigma \cdot \sqrt{\frac{\sum c_i^2}{n}}} \tag{1}$$

123 which reduces to the classical t-test for two-independent samples for the special case of

124 $\boldsymbol{c} = \{1, -1\}$. The error term of the t-test $\sigma$ is the within-group pooled standard deviation

125 (Cohen, 1988). For simplicity, we assume that all groups share the same standard deviation

126 (homoschedasticity) and the same numerosity $n$ (balanced designs). The F-test associated

127 with a contrast is simply $F_{1,k(n-1)} = t_{k(n-1)}^2$.

---

[1]https://github.com/mcfanda/cpower

**Cohen's $\delta$ measures for contrasts**

129    In his seminal work on power analysis, Cohen (1988) defines several indices of effect

130 size for the comparison of two means. In the context of two-groups designs, he defines:

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{2}$$

131 When the same logic is applied to a contrast comparison, it naturally generalizes to (Bonett,

132 2009; cf. Steiger, 2004)

$$\delta_0 = \frac{\sum (c_i \cdot \mu_i)}{\sigma} \tag{3}$$

133    Steiger (2004) refers to this measure as the standardized effect size for a contrast, but

134 the term *standardized* should be qualified to avoid confusion (Kelley, 2007; see also Steiger &

135 Fouladi, 1997). The index is standardized using the within-group standard deviation, so it

136 expesses the contrast value in terms of it. However, the contrast value depends on the

137 particular choice of weights one makes, and thus its expected value is arbitrary if no

138 restriction to the weights is applied. In fact, if we rescale the contrast weights by multiplying

139 them by a constant value, the contrast value changes and so does $\delta_0$. This seems at odds

140 with the well-known fact that in contrast analysis the scale of the contrast weights is

141 immaterial: A linear trend contrast $\boldsymbol{c_1} = \{-3, -1, 1, 3\}$ tests the same hypothesis than the

142 contrast $\boldsymbol{c_2} = \{-1, -1/3, 1/3, 1\}$. However, they give a different $\delta_0$ simply because $\boldsymbol{c_1}$

143 weights are three times larger than the ones in $\boldsymbol{c_2}$. One solution is to impose contraints to

144 the contrast weights (see, for instance, Abelson & Prentice, 1997; Kelley, 2018; Steiger &

145 Fouladi, 1997), but this strategy may result cumbersome to implement and generate

146 confusion across different applications of contrast analysis. A simpler solution is to allow

147 uncostrained weigths and to scale $\delta_0$ to render it a standardized measure of effect size.

148    It turns out, crucially, that how the index is scaled leads to different interpretations of

149 the resulting $\delta$ index, with different degrees of inferential testability and comparability. We

150 now consider two general methods to scale $\delta_0$ and discuss their properties.

151 **Normalized contrast effect size measure**

152     The first method to scale the unscaled index is to divide the contrast value by the

153 square root of the contrast weights sum of squares (Liu, 2013; Steiger & Fouladi, 1997, Lai

154 and Kelley (2012); Wahlsten, 1991). The population effect size index is:

$$\delta_z = \frac{\sum (c_i \cdot \mu_i)}{\sigma \sqrt{\sum c_i^2}} \tag{4}$$

155 Although the method employs a normalization of the weights (Liu, 2014), we refer to it as

156 the z-method and to the effect size measure as $\delta_z$, because it essentially entails as sort of

157 standardization of the contrast weights[2]. In general, it is useful to express the normalized

158 effect size index in terms of the unscaled index $\delta_0$: If we set $z = 1/\sqrt{\sum c_i^2}$, we obtain:

$$\delta_z = z \cdot \delta_0 \tag{5}$$

159     This measure has several advantages over other scaling methods but it also yields some

160 counterintuive results, both in terms of interpretability and comparability.

161     **Interpretation.**   The index expresses the contrast value in terms of the contrast

162 standard deviation. This means that it does not convey the effect in terms of the

163 within-group pooled standard deviation, as the classical Cohen's d does. In fact, when

164 applied to the special case of comparing two groups means, it yields a different value than

165 Cohen's original index. In particular, if we denote[3] as $\delta_{02}$ the original Cohen's index for

166 comparing two group means, we have that:

$$\delta_{z2} = \frac{\delta_{02}}{\sqrt{2}} \tag{6}$$

---

[2]Normalization means to divide the contrasts by $\sqrt{\sum c_i^2}$, which is the norm of **c**.

[3]Throughout the article we often refer to effect size indices computed with different scaling methods for different number of means. To clarify the notation, we use the first subscript of the index to indicate the scaling method and the second subscript to indicate the number of means in the design. For example, $\delta_{z3}$ is the effect size computed with the z-method for a contrast spanning across three means. We leave out the second subscript for general formulations of the index.

167 Its value is equivalent to the effect size one would obtain if the contrast value $\delta_0$ was tested

168 against zero in one-sample design with $n$ participants, where $n$ is the size of each group in

169 the two-group design (cf. Cohen, 1988, p. 46, index $d_3$). It is important to notice that when

170 a researcher is reporting $\delta_z$, its value and its interpretation will not correspond to classical

171 Cohen's d in a two-group design, nor to other variants of the d-index. Consider, for instance,

172 the normalized effect size for $k$ means taken two at a time discussed in Grissom and Kim

173 (2005) (p. 127-128). The authors present an unscaled $\delta$ measure comparing two means out of

174 $k$ means which is substantially equivalent to our $\delta_0$ for a contrast $\boldsymbol{c} = \{1, -1, 0, ..., 0\}$. If a

175 researcher computes Grissom and Kim (2005) effect size and compares it with the $\delta_z$ effect

176 size, the results will be different by a factor of $z$. Thus, whereas $\delta_z$ is a legitimate effect size

177 for contrasts, it does not represent a natural generalization of Cohen's d measure.

178 **Comparability.** One of the consequences of the definition of the delta index

179 computed with the z-method is that it may give surprising results when one is trying to

180 translate an observed value to a planned research with a different design, an operation often

181 done in power analysis. Consider, for instance, a case (cf. Cohen, 1988, p. 227) in which a

182 researcher has observed a mean difference of 2.5 in a two-groups design, with $\boldsymbol{\mu_2} = \{4.5, 2\}$,

183 such that $\delta_{z2} = 2.5/(\sigma\sqrt{2})$. She wishes to test the same difference in a three groups design in

184 which she expects two groups to show the same mean of the first original group, and the

185 third group as the second original group: that is, $\boldsymbol{\mu_3} = \{4.5, 4.5, 2\}$. She lays out the

186 contrast weights $\boldsymbol{c} = \{1/2, 1/2, -1\}$. The expected contrast value (means comparison) is

187 clearly equavalent in the two designs, namely $c\mu = 2.5$, but the expected effect size measure

188 will be different. In the three groups design, she will obtain $\delta_{z3} = 2.5/(\sigma\sqrt{1.5})$. Thus, the

189 comparison of the same mean values yields two different effect size outcomes. Obviously, the

190 actual size of the contrast coefficients are immaterial here, because the index is normalized.

191 The reason of the discrepancy in the example is that the standard deviation of the contrast

192 decreases with increasing number of groups, even though the pooled within-group standard

193 deviation is the same across designs. Cohen (1988, pp. 276–278) discusses several cases in

194  which he envisaged comparing designs with different number of groups sharing the same $\delta_0$.

195  In all these comparisons $\delta_z$ yields, somehow counterintuively, different effect sizes.

196      **Inferential testability.**   The great advantage of the normalized contrast index $\delta_z$ is

197  that it is easily associated with the inferential tests commonly used in testing the contrast

198  hypothesis and to compute power parameters in power analysis. In fact, $\delta_z$ is strictly related

199  to the t-test testing the contrast hypothesis. In particular:

$$E(t_{k(n-1)}) = \sqrt{n} \cdot \delta_z \tag{7}$$

200  which is the non-centrality parameter of the t-distribution used to compute the power

201  function of the t-test associated with the contrast being examined (Liu, 2014; Steiger, 2004).

202  This makes the normalization of the contrast a very handy scaling method when the main

203  concern of the researcher is to compute power and power analysis is conducted with software

204  that allows to specify the non-centrality parameter.

205      Furthermore, $\delta_z$ is strictly related also with two other effect size measures often used in

206  power analysis and meta-analysis, the $f$ and the $\eta^2$ (Cohen, 1988; Liu, 2013):

$$f = \frac{1}{\sqrt{k}} \cdot \delta_z \tag{8}$$

207      and

$$\eta^2 = \frac{\delta_z^2}{\delta_z^2 + k} \tag{9}$$

208  **Scaled effect size measure**

209      A different method of scaling the constrat effect size measure which guarantees better

210  interpretability and comparability can be suggested. Let's $g = \frac{2}{\sum_i |c_i|}$, where $|c_i|$ indicates the

211  absolute value of $c_i$, then

$$\delta_g = g \cdot \delta_0 = \frac{2}{\sum |c_i|} \cdot \frac{\sum_i c_i \cdot \mu_i}{\sigma} \tag{10}$$

212   To be able to distinguish different effect size conceptualizations, we shall denote this measure

213   of contrast effect size as $\delta_g$ and refer to it as computed with the g-method, for short. This

214   method of scaling is equivalent to constraining the contrast weights such that $\sum |c_i| = 2$, as

215   suggested by some authors (Kelley, 2018; Lai & Kelley, 2012). Thus, effect size indices scaled

216   with the g-method are in the same scale of indices computed with such a constraint.

217        This scaling guarantees several advantages over the normalized effect size $\delta_z$, although

218   it is not devoid from nuisances.

219   **Interpretation.**   The effect size computed with the g-method expresses the contrast

220   value in terms of the pooled within-group standard deviation, but it constraints the contrast

221   value to a comparable scale, the scale of the Cohen's d. Thus, it keeps the within-group

222   standard deviation as the standardization scale, but makes the contrasts comparable across

223   designs. In fact, in the special case of comparing two means, i.e. $\boldsymbol{c} = \{1, -1\}$, it yields

224   exactly the same value of the Cohen's d.

$$\delta_g = \frac{2}{2} \cdot \frac{\mu_1 - \mu_2}{\sigma} = \delta_0 \tag{11}$$

225   Thus, it can be interpreted as the most used effect size measure in the literature. The

226   particular choice of contrast weights does not matter, because the scaling compensates for

227   the arbitrary parametrization of the contrast. In the two-group design, for instance, if one

228   uses a contrast $\boldsymbol{c} = \{3, -3\}$, one obtains:

$$\delta_g = \frac{2}{6} \cdot \frac{3\mu_1 - 3\mu_2}{\sigma} = \frac{3}{3}\delta_0 = \delta_0 \tag{12}$$

229   For larger designs with $k > 2$, the index keeps the expected meaning of Cohen's d, such that

230   one can say that, for any number of groups $k$, a contrast with a given $\delta_{gk} = d$ shows the

231   same effect size than two groups with a standardized mean difference of $d$.

232        In general terms, the g-method explicitly deals with the correct coding of the contrast

233   weigths. Several authors have discussed standardized measures of contrast effect size,

234   without explicitly defining how to transform an arbitrarily coded contrast into a properly

235   coded one. Either in discussing examples (Kelley, 2007; Steiger, 2004), or in the

236  implementation of software (Kelley, 2018, p. 30), it is often implicitily assumed that the

237  contrast weights posses some characteristic which makes the standardized contrast

238  interpretable as a Cohen's d. Other authors suggested similar scaling but applied only to

239  limited cases, such as contrasts applied to two-group designs (Abelson & Prentice, 1997).

240  The g-method makes the transformation explicit, such that it can be employed with any

241  arbitrary set of contrast weights.

242       **Comparability.**   Consider, again, the case (cf. Cohen, 1988, p. 227) in which a

243  researcher has observed a mean difference of 2.5 in a two-groups design, with $\boldsymbol{\mu_2} = \{4.5, 2\}$,

244  such that $\delta_{02} = 2.5/\sigma$. In a three-group design he expects $\boldsymbol{\mu_3} = \{4.5, 4.5, 2\}$, and he

245  evaluates the constrast $\boldsymbol{c} = \{1/2, 1/2, -1\}$. This is the case where the expected contrast

246  value is equavalent in the two designs, namely $c\mu_3 = 2.5$. In virtue of the $\delta_g$ charateristics, in

247  the three-group design the researcher will obtain

$$\delta_{g3} = \frac{2}{2} \cdot \frac{2.5}{\sigma} = \frac{2.5}{\sigma}$$

248  which is exactly what he obtains in the two groups design. Thus, the comparison of the same

249  mean values yields identical effect size measures. One can also notice that the effect size

250  quantity will be the same independently of the number of groups, $k$, provided the expected

251  means are comparable. In particular, if one group has $\mu_1 = d$ and all other groups share the

252  same mean $\mu_{2..k} = 0$, and the comparison is tested with the contrast

253  $\boldsymbol{c} = \{(k-1), -1_2, -1_3, .., -1_k\}$, the contrast value is $c\mu = (k-1) \cdot d$. The g-method yields,

254  after simple algebra, $g = 1/(k-1)$, and thus $\delta_{gk} = \delta_{02}$, independently of the number of

255  groups $k$.

256       It is easy to verify that in all the cases discussed by Cohen (1988, pp. 276–278) in

257  which he envisaged comparing designs with different number of groups sharing the same $\delta_0$,

258  $\delta_g$ gives the same effect size quantity.

259       **Inferential testability.**   One drawback of the scaled effect size measure is that it

260  does not correspond directly to the inferential tests used to test hypotheses about the

261 contrast or to compute the associated power function. Its correspondence, however, can be

262 obtained by simple transformations. One needs to transform $\delta_g$ into a $\delta_z$ and exploit the

263 latter index inferential testability properties. Thus, it is useful to relate the two indices:

$$\delta_z = \frac{z}{g} \cdot \delta_g = \frac{\sum_i |c_i|}{2\sqrt{\sum c_i^2}} \cdot \delta_g \tag{13}$$

264 It should be noted that the ratio $z/g$ is usually very easy to compute. For an interaction

265 contrast in a 2-by-2 design, $\boldsymbol{c} = \{1, -1 - 1, 1\}$, for instance, one can mentally verify that it is

266 4 divided by 4, that is 1. From equation (13) it follows that the expected value of the t-test

267 associated with $\delta_g$ is:

$$E(t_{k(n-1)}) = \sqrt{n} \cdot \frac{z}{g} \delta_g \tag{14}$$

268 and the index is related with other effect size indices as follows:

$$f = \frac{1}{\sqrt{k}} \cdot \frac{z}{g} \cdot \delta_g \tag{15}$$

$$\eta_p^2 = \frac{\delta_g^2}{\delta_g^2 + (\frac{g^2}{z^2} \cdot k)} \tag{16}$$

269 ## Sample Estimation

270    As for the classical Cohen's d, estimating the population effect size using sample data

271 may take different routes depending on the charateristics of the available data (Grissom &

272 Kim, 2005). For $\delta$-like measures, those routes differ in the way the within-group pooled

273 standard deviation $\sigma$ is estimated. It is important to clarify, however, that the way $\sigma$ is

274 estimated in the sample may change the numerical value of the effect size indices discussed

275 here, but it does not alter their properties. In contrast analysis, the contrast is almost always

276 estimated within the framework of the ANOVA, and the standard deviation $\sigma$ is estimated

277 as the square root of the mean square error of the ANOVA, namely (Howell, 2012, p. 204,

278 p. 380; Kelley, 2007, Lai and Kelley (2012); Rosenthal et al., 2000, p. 41; Steiger, 2004):

$$s_p = \sqrt{\frac{\sum_i (n_i - 1)s_i^2}{N - k}} \tag{17}$$

279  which simplifies to $\sqrt{s^2}$ when groups have the same numerosity $n$ and the same variance $s^2$.

280  This choice of estimate of $\sigma$ makes the effect size computed for the contrast equivalent to the

281  Hedges's estimation of the standardized mean difference (Hedges & Olkin, 1985), often

282  reffered to as *Hedges's g*. Other forms of estimation of $\sigma$ are plausible, that can accomodate

283  heteroskedasticity and different numerosity in the group, and reduce sampling bias (Grissom

284  & Kim, 2005; Lai & Kelley, 2012).

285      Once $s_p$ is available, and the population means are estimated with the sample means

286  $\boldsymbol{m} = \{m_i\}$, for any given contrast $\boldsymbol{c} = \{c_i\}$, one can compute the scaled effect size indices.

287  As regards estimating $\delta_z$, one can lay out a contrast with arbitrarily scaled weights and

288  compute:

$$d_z = \frac{\sum (c_i \cdot m_i)}{s_p \sqrt{\sum c_i^2}} \tag{18}$$

289  Alternatively, one may first normalize the contrast weights and then compute

290  $d = \sum_i (c_i m_i)/s_p$. Obviously, the latter method works because when the weigthts are

291  normalized, $\sqrt{\sum c_i^2} = 1$, thus $\sum_i (c_i m_i)/s_p = d_z$. This also suggests that the z-method is

292  handy for users of software that provides normalize contrasts weights by default, such as

293  SPSS.

294      As regards estimating $\delta_g$, one can lay out a contrast with arbitrarily scaled weights,

295  and compute

$$d_g = 2 \cdot \frac{\sum (c_i \cdot m_i)}{s_p \sum |c_i|} \tag{19}$$

296  or pick contrast weights that show $g = 1$. This can be achieved by constraining the weights

297  such that $\sum |c_i| = 2$ (Lai & Kelley, 2012). Notice that there are several circumstances where

298  choosing contrast weights with $g = 1$ is extremely easy. For instance, any contrast which

299  assigns either 1 or -1 to the means, such as main effects and interactions contrasts, has $g = 1$

if weights equal to 1 and -1 are replaced with $2/k$ and $-2/k$, respectively. Furthermore, any contrast featuring fractional weigths in which the positive weights sum up to 1 guarantees $g = 1$.

In meta-analysis and power analysis raw data are often not available, thus the sample estimation can be achieved starting from the inferential test associated with the effect (Glass et al., 1981). When the t-test is available, one can estimate the contrast effect size measures as (cf. Kelley, 2007; Steiger, 2004):

$$d_z = \frac{t_{k(n-1)}}{\sqrt{n}} \tag{20}$$

and

$$d_g = \frac{t_{k(n-1)}}{\sqrt{n}} \cdot \frac{g}{z} \tag{21}$$

assuming[4] that each group has the same numerosity $n$ . When the F-test is available, one just recalls that $t_{k(n-1)} = \sqrt{F_{1,k(n-1)}}$ and derives the effect size indices accordingly.

A special mention should be done for Rosenthal et al. (2000) contrast effect sizes based on the correlation metric. Rosenthal et al. (2000) define three effect size indices, one of which, the $r_{contrast}$, can be related with the d-like measures presented here. The $r_{contrast}$, when squared, represents the sample estimate of the $\eta_p^2$ associated with the contrast, defined in (9) and (16). Thus, if a researcher needs to translate $r_{contrast}$ into the proposed d-like measures, this relation can be used to find the appropriated translation formulas. However, it should be noted that the $r_{contrast}$ is defined for the sample, not the population, thus the sample size should be taken into the account (Rosnow, Rosenthal, & Rubin, 2000, EQ 9). After some algebra, one can derive the following transformation formulas:

$$\delta_z = \sqrt{k \cdot \frac{n-1}{n}} \cdot \frac{r_{contrast}}{\sqrt{1 - r_{contrast}^2}} \tag{22}$$

and

$$\delta_g = \frac{g}{z} \sqrt{k \cdot \frac{n-1}{n}} \cdot \frac{r_{contrast}}{\sqrt{1 - r_{contrast}^2}} \tag{23}$$

---

[4]Notice that for $k = 2$, $d_{g2} = t_{(N-2)} \cdot \sqrt{\frac{2}{n}}$ as expected (Glass, McGaw, & Smith, 1981, p. 126, EQ 5.37; Rosenthal et al, 2000, p. 12, EQ 2.10)

320        The other two effect size indices defined by Rosenthal et al. (2000), namely $r_{effectsize}$

321    and $r_{alert}$, are not directly translatable into a d-like measure because they embed also

322    variance which is not part of the within-group variance or the variance explained by the

323    contrast, and thus they are not comparable with the proposed indices.

<div align="center">

## Power Analysis

</div>

325        A contrast hypothesis is usually tested with the t-test (Rosenthal et al., 2000; Steiger,

326    2004). Power analysis software usually provides a way to compute the power parameters ($\beta$,

327    required $n$ per cell, etc.) of a t-test based on the expected $\delta$. The critical bit of information

328    required for power calculation that is interesting here is the non-centrality parameter $\lambda$

329    which affects the location of the t-distribution employed in computing power. For standard

330    two independent samples t-test the parameter is (Cohen, 1988):

$$\lambda_2 = \sqrt{\frac{n}{2}} \cdot \delta \tag{24}$$

331        From the perspective of contrast analysis, $\lambda_2$ is a special case of a more general

332    non-centrality parameter of the t-test with $k(n-1)$ degrees of freedom associated with a

333    contrast $\boldsymbol{c}$ (Liu, 2014; Steiger, 2004):

$$\lambda_k = \sqrt{\frac{n}{\sum_i c_i^2}} \cdot \delta_0 \tag{25}$$

334    which equates $\lambda_2$ when $\boldsymbol{c} = \{1, -1\}$ as in the two independent samples t-test. Because

335    different scaling methods produce indices with different relations with $\delta_0$, they require

336    different transformations to obtain the correct non-centrality parameter.

337        When data are available for the computation of the non-centrality parameter one can

338    simply estimate it as:

$$\hat{\lambda}_k = \sqrt{\frac{n}{\sum_i c_i^2}} \cdot d_0 \tag{26}$$

339 When only a scaled $d$ is available, the non-centrality parameter needs to be scaled back to

340 $\sqrt{n/\sum_i c_i^2} \cdot d_0$ accordingly to the used scaling method. When $d$ is an estimation of $\delta_z$,

$$\hat{\lambda}_k = \sqrt{n} \cdot d_z \tag{27}$$

341 When $d$ is an estimation of $\delta_g$,

$$\hat{\lambda}_k = \sqrt{n} \cdot \frac{z}{g} \cdot d_g = \sqrt{n} \cdot \frac{\sum |c_i|}{2 \cdot \sqrt{\sum_i c_i^2}} \cdot d_g \tag{28}$$

342 In the last equation $g$ is simply a constant that multiplies $d_0$ to obtain a scaled $d_g$. Thus,

343 whatever scaling $w$ one is using such that $\delta_w = \delta_0 \cdot w$, we have:

$$\lambda_k = \sqrt{n} \cdot \frac{z}{w} \cdot \delta_w \tag{29}$$

344 It may come as a surprise that different effect size quantities share the same non-centrality

345 parameter and thus the same power function. However, this is always the case for effect size

346 measures that refer to the same population parameter. In the linear model, for instance, the

347 $B$ and *beta* coefficients have very different scales, but they refer to the same population

348 effect. Coherently, they lead to the same power function in power analysis.

349     A contrast hypothesis can also be tested with the F-test (Rosenthal et al., 2000).

350 Power analysis software usually provides a way to compute the power parameters based on

351 the F-test. The non-centrality parameter of the F-test is given by (Cohen, 1988, p. 481;

352 Steiger, 2004):

$$\hat{\lambda}_F = k \cdot n \cdot f^2 \tag{30}$$

353 thus, one simply transforms $\delta_z$ or $\delta_g$ into $f$ and then compute the power of the F-test

354 associated with 1 and $k(n-1)$ degrees of freedom.

355 **Software usage**

356     **G\*Power.**   G\*Power provides power functions for "generic t-test" which allows to

357 input $\lambda$ and *df* and returns the power. One can compute the non-centrality parameter as

358 shown above and input it in the software. Unfortunately, the "generic t-test" function of the

359 software does not allow to estimate the required N, an operation often useful for users.

360 However, in G*Power one can compute all power parameters of a contrast using the F-test.

361 Under "ANOVA: fixed effect, special, main effects and interactions" it is possible to specify k

362 (number of groups), $df = k(n-1)$ and the effect size $f$. The correct $f$ can be computed from

363 $d_z$ and $d_g$ using equation (8) and (15).

364        **R.**    To the best of our knowledge, R power functions commonly used in t-test power

365 analysis do not allow to accomodate for contrasts with arbitrary weights. This is mainly due

366 to the computation of $df$ and $\lambda$ that are tailored either to one-sample t-test, where

367 $\lambda = \sqrt{N} \cdot \delta$ and $df = N - 1$, or to two-samples t-test, where $\lambda = \sqrt{n} \cdot \delta$ and $df = N - 2$

368        Thus, we have written a simple power function that computes the power parameters

369 for contrasts based on estimated $\delta$ scaled in arbitrary ways, with shortcuts for the g-method

370 and z-method. They are included in the `cpower` R package.

## Confidence intervals

372        Confidence intervals for the estimated effect size indices can be computed using the

373 *non-centrality interval estimation* method defined by Steiger and Fouladi (1997; se also

374 Steiger, 2004). Their method entails to compute the confidence interval for the

375 non-centrality parameter of the distribution associated with the effect size index, and then

376 trasform the limits of that interval to the scale of the effect size.

377        When $d$ is an estimation of $\delta_z$, recall that the non-centrality parameter is equal to the

378 observed t-test:

$$\hat{\lambda}_t = t_{k(n-1)} = \sqrt{n} \cdot d_z \tag{31}$$

379 The lower ($\hat{\lambda}_l$) and the upper ($\hat{\lambda}_u$) limit of the $100(1 - \alpha)\%$ confidence interval can be

380 established by finding two noncentral distributions for which the value $t$ represents the

381 $100(1 - \alpha/2)$-th and $100(\alpha/2)$-th percentile, respectively. This method yields the confidence

382    interval in the scale of the non-centrality parameter:

$$Pr\left[\hat{\lambda}_l \le \hat{\lambda}_t \le \hat{\lambda}_u\right] = 1 - \alpha \tag{32}$$

383    When the interval is computed, one transforms the limits to scale them to the effect size

384    scale (cf. Lai & Kelley, 2012):

$$Pr\left[\frac{\hat{\lambda}_l}{\sqrt{n}} \le d_z \le \frac{\hat{\lambda}_u}{\sqrt{n}}\right] = 1 - \alpha \tag{33}$$

385        For the g-method effect size, one proceeds in the same way, but uses a different

386    transformation, namely:

$$Pr\left[\frac{2}{\sum |c_i|} \cdot \sqrt{\frac{\sum c_i^2}{n}} \cdot \hat{\lambda}_l \le d_g \le \frac{2}{\sum |c_i|} \cdot \sqrt{\frac{\sum c_i^2}{n}} \cdot \hat{\lambda}_l\right] = 1 - \alpha \tag{34}$$

387    It is easy to verify that for the two groups design, the confidence interval around $d_{g2}$

388    simplifies to:

$$Pr\left[\sqrt{\frac{2}{n}} \cdot \hat{\lambda}_l \le d_{g2} \le \sqrt{\frac{2}{n}} \cdot \hat{\lambda}_u\right] = 1 - \alpha \tag{35}$$

389    and reduces to the Cohen's d confidence interval (Kelley, 2007), as expected. It is useful to

390    reiterate that the confidence interval for $d_z$ does not correspond to Cohen's d interval, simply

391    because the two indices are generally different. It should be noted, however, that when the

392    confidence interval is used to reject the null-hypothesis, both $d_g$ and $d_z$ lead to the same

393    conclusion. In fact, the scaling of the noncentral parameter does not change the sign of the

394    limits, thus if the interval around $d_g$ contains zero, so does the interval around $d_z$, and

395    viceversa. The R package `cpower` accompaning this contribution provides functions to

396    compute confidence intervals for any contrast.


397                                           **Examples**


398        We now consider two examples to outline the methods discussed in this contribution in

399    practical terms. We employ the R package `cpower` and make reference also to other software

400  which can give the same results (See Appendix 1 for actual commands used in the examples).

401  In the first example we focus on computation of the effect size indices and their confidence

402  intervals when data are available. Imagine a four groups design, with 30 participants in each

403  group and each group representing an increasing level of a manipulated stimulus. The

404  response to the stimulus has been recorded for each participant on a continuous scale. Data

405  are reported in Table 1.

|  | grp1 | grp2 | grp3 | grp4 |
| --- | --- | --- | --- | --- |
| Mean | 8.00 | 16.00 | 18.00 | 19.00 |
| SD | 7.00 | 7.50 | 7.40 | 7.00 |
| Linear | -3.00 | -1.00 | 1.00 | 3.00 |
| Quadratic | -1.00 | 1.00 | 1.00 | -1.00 |

Table 1

*Means, standard deviations and contrast weights for the four groups example*

406      The pooled standard deviation is of 7.23. We wish to estimate and quantify the linear

407  and the quadratic trend of the four means. Thus, we lay out two sets of weights as described

408  in Table 1.

409      As regards the linear trend, we obtain a statistically significant t-test, $t(116) = 5.93$,

410  p.<.001, with a $d_g = 1.21$ and 95% confidence interval with limits [0.78,1.64]. We can then

411  say that the linear trend shows a strong effect, equivalent to a Cohen's d of 1.21. More

412  precisely, the average increment in mean response is 1.21 standard deviations across levels of

413  stimulus. The standardized effect size is $d_z = 1.08$, with confidence limits [0.70,1.46]. As

414  expected $d_z$ is smaller than $d_g$, and its confidence interval slightly narrower.

415      As regards the quadratic trend, we obtain a statistically significant t-test, $t(116) =$

416  2.65, p.=0.01, with a $d_g = 0.48$ and 95% confidence interval with limits [0.12,0.85]. The

417  quadratic trend shows a medium effect, equivalent to a Cohen's d of 0.48. The quadratic

418  trend is much smaller than the linear trend, thus the increment of response across stimulus

419   levels is more important than the curvature that the trend shows. The normalized effect size

420   is $d_z = 0.48$, with confidence limits [0.12,0.85]. It is not surprising that for the quadratic

421   trend we obtain $d_g = d_z$. In fact, in the quadratic $\sum |c_i^2| = 2 * \sqrt{\sum c_i^2}$ and thus $g = z$.

422          A second interesting example regards the computation of effect size indices in power

423   analysis. The example is inspired by Wahlsten (1991) treatment of 2x2 designs. A researcher

424   observes in a two-group design a factor A with means $\boldsymbol{m} = \{10, 7\}$ and $s_p = 4$. She wishes to

425   run a 2 A x 2 B design where A is the same factor as in the two-group design, and B is a

426   moderator. Let $\mu_{AB}$ be the expected mean for $A = \{1, 2\}$ and $B = \{1, 2\}$. She expects to

427   replicate the two-group effect of A in condition $B = 1$, and absence of effect of A in

428   condition $B = 2$. That is, $\{\mu_{11} = 10, \mu_{21} = 7, \mu_{12} = 7, \mu_{22} = 7\}$. Figure 1 represents the
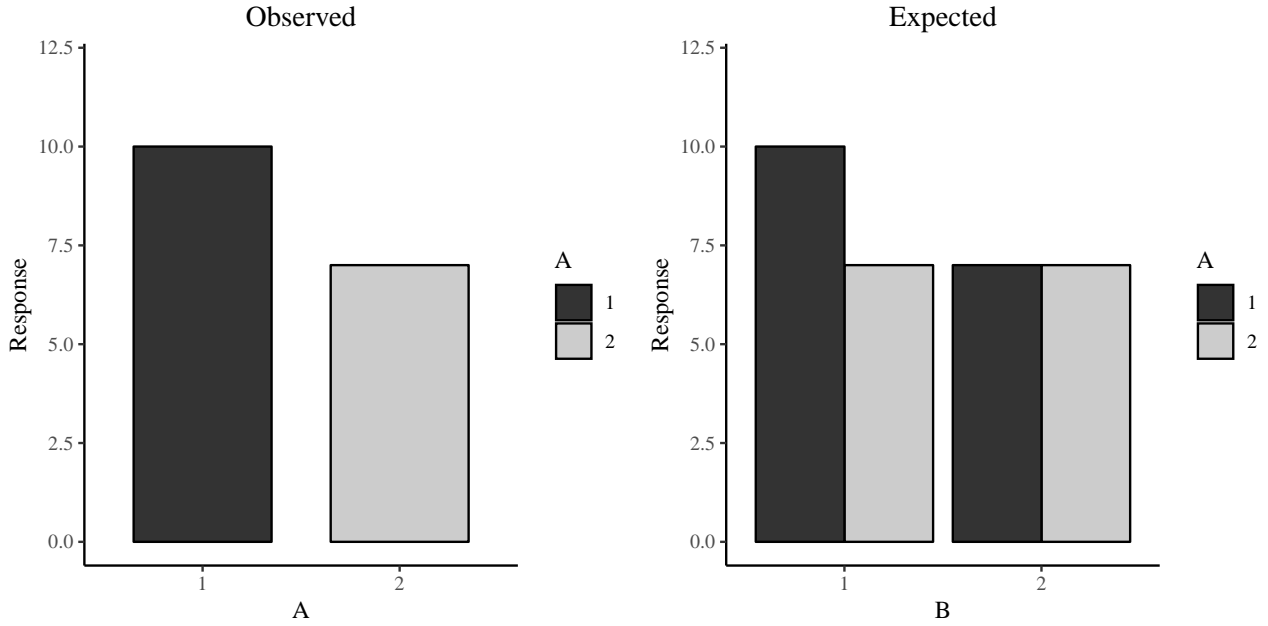
429   observed and the expected results.



*Figure 1*. Observed and expected means in the example with two designs

430          The researcher wishes to estimate the expected $\delta_g$ for the interaction A X B, and its

431   power parameters. For the 2-groups design, the observed effect size is:

$$\delta_{g2} = \frac{2}{2} \cdot \frac{10 - 7}{4} = \frac{3}{4}$$

432   The interaction effect contrast can be written as $\boldsymbol{c} = \{1, -1, -1, 1\}$. Computing the expected

effect size using g-method yields:

$$\delta_{g4} = \frac{2}{4} \cdot \frac{10 - 7 - 7 + 7}{4} = \frac{3}{8}$$

Thus, the effect size of the interaction is half the effect size of the effect obtained in the two-group design. Notice that for the interaction it would be tempting to compute the effect size without a scaling method, but the result will be wrong.

Based on the estimation of $\delta_g$, we can now compute the sample size required to achieve .80 power in the 2X2 design. Using `power.contrast.t` function from `cpower` R package, we obtain n=56, for a total sample size of N=224. Coherently with the effect size measures, the same level of power can be achieved in the original two groups design with n=29, for a total sample size of N=58.

We can achieve the same results by employing G*Power. Transforming $d_g$ in $f$ yields an $f = .188$ . The "ANOVA: fixed effect, special, main effects and interactions" function of G*power, with 4 groups indicates a total sample size of 225, which is in line with the results obtained with `cpower` R package.

## Conclusions

Planned comparisons and in general contrast analysis can benefit from using standardized effect size measures that can be interpreted as standardized mean differences, and compared with the Cohen's d, one of the most widely used effect size index. We have discussed two methods to obtain interpretable, comparable, and testable effect size indices within this class of measures.

The two methods are characterized by a transparent logic. By keeping separate contrast weights and their scaling, a researcher can decide upon the appropriate weights for a contrast following established rules and properties, most notably that the the scale of the contrast weights is immaterial. The scaling of the effect size index, instead, reflects what kind of properties of the index one wishes to select.

⁴⁵⁷        Balancing advantages and disavantages, method-g $\delta_g$ index should be preferred over

⁴⁵⁸ other scaling methods of the $\delta$-like measures of effect size. It keeps the same scale of Cohen's

⁴⁵⁹ d, and thus can be compared with published results. It also nicely generalizes the Cohen's d

⁴⁶⁰ logic to multi-groups designs. An important property in fact is its comparability across

⁴⁶¹ different designs. This property allows for consistency in interpreting the size of the effect as

⁴⁶² well as it is lined up with the researcher's intuition who would expect an effect size index to

⁴⁶³ provide the same value when the difference between the mean values are the same, regardless

⁴⁶⁴ of how many groups are involved.

⁴⁶⁵        Nonetheless, the z-method scales can also be considered. This method of normalizing

⁴⁶⁶ the contrasts weights has gathered popularity and thus it can be often the case that

⁴⁶⁷ published research provides, although implicitly, this form of effect indices. As long as the

⁴⁶⁸ researcher uses this method consistently across designs, the method is useful for quick

⁴⁶⁹ calculations of power parameters. Indeed, in meta-analysis both scaling methods can be

⁴⁷⁰ used, provided that only one of them is used in the same analysis. When authors of

⁴⁷¹ published research have reported effect sizes with different scaling methods, the meta-analyst

⁴⁷² should convert them using the same scale. When published research includes effects reported

⁴⁷³ as Cohen's d (and their variants), using the g-method is advisable, because it retains the

⁴⁷⁴ same scale of the two-group index.

## Appendix: Example commands

This section contains the R commands used to produce the examples.

**Example 1**

```
library(cpower)
n<-30
m=c(8,16,18,19)
sp<-7.23
k<-length(m)

## linear trend ##
linear<-c(-3,-1,1,3)
dg_linear<-d.contr(linear,means = m,sd=sp,scale = "g")
test.contr(linear,d=dg_linear,n=n,scale = "g")
ci.contr(linear,dg_linear,n = n)
dz_linear<-d.contr(linear,means = m,sd=sp,scale = "z")
ci.contr(linear,dlz,n = n,scale = "z")

## quadratic trend ##

quad<-c(-1,1,1,-1)
dg_quad<- d.contr(quad,means = m,sd=sp,scale = "g")
test.contr(quad,d=dg_quad,n=n,scale = "g")
ci.contr(quad,dg_quad,n = n)

dz_quad<- d.contr(quad,means = m,sd=sp,scale = "z")
ci.contr(quad,dz_quad,n = n,scale = "z")
```

501

## Example 2

```
m2<-c(10,7)

c2<-c(1,-1)

m4=c(10,7,7,7)

c4<-c(1,-1,-1,1)

sp<-4

k<-length(m)


## effect size computation

dg2<-d.contr(c2,means = m2,sd=sp,scale = "g")

dg4<-d.contr(c4,means = m4,sd=sp,scale = "g")


##  required n per cell  ##

power.contrast.t(c4,dg4,power = .80,scale = "g")

power.contrast.t(c2,dg2,power = .80, scale="g")


## transform dg in f  ##

f<-f.contr.d(c4,dg4,scale = "g")
```

## References

Abelson, R. P., & Prentice, D. A. (1997). Contrast tests of interaction hypothesis. *Psychological Methods*, *2*(4), 315.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . others. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

Bakker, M., Dijk, A. van, & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Bonett, D. G. (2009). Estimating standardized linear contrasts of means with desired precision. *Psychological Methods*, *14*(1), 1.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences . hilsdale. *NJ: Lawrence Earlbaum Associates*, *2*.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Furr, R. M. (2004). Interpreting effect sizes in contrast analysis. *Understanding Statistics*, *3*(1), 1–25.

Furr, R. M., & Rosenthal, R. (2003). Evaluating theories efficiently: The nuts and bolts of contrast analysis. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *2*(1), 33–67.

Glass, G. V., Smith, M. L., & McGaw, B. (1981). *Meta-analysis in social research.* Sage Publications, Incorporated.

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

Lawrence Erlbaum Associates Publishers.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando FL: Academic Press.

Henson, R. K., & Smith, A. D. (2000). State of the art in statistical significance and effect size reporting: A review of the apa task force report and current trends. *Journal of Research & Development in Education*.

Howell, D. C. (2012). *Statistical methods for psychology*. Cengage Learning.

IBM Corp. (2017). *Released 2017. ibm spss statistics for windows, version 25.0*. Armonk, NY.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2011). *Data analysis: A model comparison approach*. Routledge.

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*(8), 1–24.

Kelley, K. (2018). *MBESS: Methods for the behavioral, educational, and social sciences*. Retrieved from https://cran.r-project.org/package=MBESS

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137.

Keppel, G. (1991). *Design and analysis*. NJ: Prentic Hall, Engelwood Cliffs.

Lai, K., & Kelley, K. (2012). Accuracy in parameter estimation for ancova and anova contrasts: Sample size planning via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 350–370.

Liu, X. S. (2013). *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. Routledge.

Liu, X. S. (2014). A note on statistical power in multi-site randomized trials with multiple treatments at each site. *British Journal of Mathematical and Statistical Psychology*, *67*(2), 231–247.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research:

Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective.* Routledge.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods*, *11*(4), 386–401.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105.

R Core Team. (2008). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* CUP Archive.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge University Press.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, *11*(6), 446–453.

Steiger, J. H. (2004). Beyond the f test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*(2), 164.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. E. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?: Classic edition.* Lawrence Erlbaum Associates Publishers.

Wahlsten, D. (1991). Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. *Psychological Bulletin*, *110*(3), 587.

Wiens, S., & Nilsson, M. E. (2017). Performing contrast analysis in factorial designs: From nhst to confidence intervals and beyond. *Educational and Psychological Measurement*,

*77*(4), 690–715.

Wilkinson, L. (1999). Task force on statistical inference, american psychological association, science directorate.(1999). statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.