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

Author Note

This manuscript has been submitted to XXX on July XXX 2018. Please do not cite without permission

Correspondence concerning this article should be addressed to Marcello Gallucci, Department of Psychology, University of Milano-Bicocca, Piazza dell'Ateneo Nuovo, 20126 Milan, Italy. E-mail: marcello.gallucci@unimib.it

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 indexes are in the correlation metric, which may result difficult to interpret and to generalize across different designs. Here we discuss effect size indexes 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: One method which standardizes 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 contast 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 computation 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 data collection. However, reaserchers willing to adopt a contrast analysis approach will find difficulties in identifying a clear way to estimate contrast effect sizes from the literature. Developed effect sizes in this area are expressed in a correlational metrics (Rosenthal et al., 2000). From correlation-based effect size is easy to convert to mean-based effect size indices, yet their properties are not necessarily transferable and each has advantages and disadvantages (McGrath & Meyer, 2006). Moreover, consider that contrast analysis involves focused tests of groups means, hence the intuitive effect size unit is one that should involve differences in means in its calculation, such as Cohen's d and related variants. It seems almost paradoxical then that most developments about effect size indices in contrast analysis have been done in a correlational approach.

In this article we review measures of effect sizes for contrast analysis inspired by the classical Cohens'd class of measures, with particular focus on their interpretability, comparability, and inferential testability. By *interpretability* we mean the degree by which a measure has a clear meaning, and thus a clear interpretation, when defined for different statistical effects and applications. The increased attention for effect size measures in Psychology has included a recommendation to report and interpret the magnitude of effects, both in absolute and in relative terms (Henson & Smith, 2000; Wilkinson, 1999). Interpreting an effect size index in absolute terms is much easier when it has a clear definition and an intuitive metrics. The ease of interpretation of an effect size index relatively to the published results in a research field depends on its degree of comparability. By *comparability* (Bakeman, 2005; Glass, Smith, & McGaw, 1981; Keppel, 1991; Morris & DeShon, 2002) we mean the degree by which a measure of effect size conveys the same quantity when associated with the same effect, across different designs and applications. Comparability is crucial expecially in power analysis, in which the researcher needs to guess the expected effect size in order to compute power parameters (Beta, expected N, etc.). When the expected effect size is inspired by published research, it is often the case that the planned research is not exactly equal to the published one, thus the researcher needs to port, i.e. translate, the observed effect size to the new design. A comparable and therefore portable effect size makes this operation easier. Comparability is important also in meta-analysis: When several effect size indices are gathered from the literature and are aggregated to estimate in a reliable way the population effect sizes, the indices that are aggregated should be comparable in the metrics and in the interpretation. If this is not the case, the overall estimated effect would represent a mis-specification of the population parameters.

By *inferential testability* we mean the degree by which a measure of effect size is directly associated with an inferential test used to test (null) hypotheses about the observed results or to compute power in the planning phase of the research. There are effect size indices, in fact, that can be readily associated with a statistical test, such has Cohen's d-like measures (Cohen, 1988) and correlation indices (Rosenthal et al., 2000). This makes their usage in power analysis and meta-analysis greatly faciliated. Other effect size measures are not logically associated with an inferential test, thus their applicability may be limited to specific phases of the research development. Inferential testability can be also evaluated for a set of effect size measures. It is often the case, in fact, that an effect can be quantified with different, alternative measures of size. Such a set of measures may, with varying degrees of difficulty, be reconciled with the same inferential test, showing shared inferential testability. This property allows to conduct coherent hypothesis testing and power analysis, and greatly facilitates aggregation of results in meta-analyses. To provide an example, in regression analysis an effect size can be quantified with several different measures, such as the unstandardized B coefficient, standardized β coefficient, the η^2 or the partial η^2. All these effect size measures express the same effect using different scales and emphasizing different charateristics of the effect. However, they all share the same inferential test, usually the t-test, and the same power function. In this article we present alternative definitions of effect size measures for contrast analysis, with particular emphasis on their shared inferential testability.

Finally, we also discuss practical aspects of computation and estimation of effect size indices in contrast analysis, by discussing some theoretical results and presenting dedicated software that can be used alongside popular software, such as R (R Core Team, 2008), SPSS (IBM Corp., 2017), and G*Power (Faul, Erdfelder, Lang, & Buchner, 2007). We accompany the article with a R package named `cpower`, specifically developed to execute the statistical procedures presented here. The R package `cpower` can be found at github[1].

Contrasts analysis background

A contrast is a linear combination of means whose coefficients sum up to zero, meant to estimate a particular comparison of means and test it against zero. We refer to the contrast set of coefficients as $\boldsymbol{c} = \{c_i\}$, and to the expected set of means as $\boldsymbol{\mu} = \{\mu_i\}$. The contrast coefficients (weights) are chosen such that $\sum_i c_i = 0$, with $i = \{1, .., k\}$ where k is the number of means being compared. The contrast expected value is $c\mu = \sum_i (c_i \cdot \mu_i)$. As an example, consider a simple design with two groups: the comparison of the two groups means can be carried out with a simple contrast with $\boldsymbol{c} = \{1, -1\}$, in which the contrast value is simply the expected difference between means, $c\mu = c_1\mu_1 + c_2\mu_2 = \mu_1 - \mu_2$.

A contrast defined across k means of independent groups of size n can be tested employing either an independent samples t-test or an F-test. The t-test expected value, with $k(n-1)$ degrees of freedom, is (Steiger, 2004):

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

which reduces to the classical t-test for two-independent samples for the special case of $\boldsymbol{c} = \{1, -1\}$. The error term of the t-test σ is the within-group pooled standard deviation (Cohen, 1988). For simplicity, we assume that all groups share the same standard deviation (homoschedasticity) and the same numerosity n (balanced designs). The F-test associated with a contrast is simply $F_{1,k(n-1)} = t^2_{k(n-1)}$.

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

Cohen's δ measures for contrasts

In his seminal work on power analysis, Cohen (1988) defines several indices of effect size for the comparison of two means. In the context of two-groups designs, he defines:

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

When the same logic is applied to a contrast comparison, it naturally generalizes to (Bonett, 2009; cf. Steiger, 2004)

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

Steiger (2004) refers to this measure as the standardized effect size for a contrast, but the term *standardized* should be qualified to avoid confusion (Kelley, 2007; see also Steiger & Fouladi, 1997). The index is standardized using the within-group standard deviation, so it expesses the contrast value in terms of it. However, the contrast value depends on the particular choice of weights one makes, and thus its expected value is arbitrary if no restriction to the weights is applied. In fact, if we rescale the contrast weights by multiplying them by a constant value, the contrast value changes and so does δ_0. This seems at odds with the well-known fact that in contrast analysis the scale of the contrast weights is immaterial: A linear trend contrast $\boldsymbol{c_1} = \{-3, -1, 1, 3\}$ tests the same hypothesis than the contrast $\boldsymbol{c_2} = \{-1, -1/3, 1/3, 1\}$. However, they give a different δ_0 simply because $\boldsymbol{c_1}$ weights are three times larger than the ones in $\boldsymbol{c_2}$. One solution is to impose contraints to the contrast weights (see, for instance, Abelson & Prentice, 1997; Kelley, 2018), but this strategy may result cumbersome to implement and generate confusion across different applications of contrast analysis. A simpler solution is to allow uncostrained weigths and to scale δ_0 to render it a standardized measure of effect size.

It turns out, crucially, that how the index is scaled leads to different interpretations of the resulting δ index, with different degrees of inferential testability and comparability. We now consider two general methods to scale δ_0 and discuss their properties.

Standardized contrast effect size measure

The first method to scale the unscaled index is to divide the contrast value by the standard deviation of the contrast weights (Liu, 2013; Steiger & Fouladi, 1997, Lai and Kelley (2012); Wahlsten, 1991). The population effect size index is:

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

We can refer to this scaling method as the z-method and to the effect size measure as δ_z, because it essentially entails to standardize the contrast weights. In general, it is useful to express the standardized effect size index in terms of the unscaled index δ_0: If we set $z = 1/\sqrt{\sum c_i^2}$, we obtain:

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

This measure has several advantages over other scaling methods but it also yields some counterintuive results, both in terms of interpretability and comparability.

Interpretation. The index expresses the contrast value in terms of the contrast standard deviation. This means that it does not convey the effect in terms of the within-group pooled standard deviation, as the classical Cohen's d does. In fact, when applied to the special case of comparing two groups means, it yields a different value than Cohen's original index. In particular, if we denote[2] as δ_{02} the original Cohen's index for comparing two group means, we have that:

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

Its value is equivalent to the effect size one would obtain if the contrast value δ_0 was tested against zero in one-sample design with n participants, where n is the size of each group in

[2] 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, δ_{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.

the two-group design (cf. Cohen, 1988, p. 46, index d_3). It is important to notice that when a researcher is reporting δ_z, its value and its interpretation will not correspond to classical Cohen's d in a two-group design, nor to other variants of the d-index. Consider, for instance, the standardized effect size for k means taken two at a time discussed in Grissom and Kim (2005) (p. 127-128). The authors present an unscaled δ measure comparing two means out of k means which is substantially equivalent to our δ_0 for a contrast $\boldsymbol{c} = \{1, -1, 0, ..., 0\}$. If a researcher computes Grissom and Kim (2005) effect size and compares it with the δ_z effect size, the results will be different by a factor of z. Thus, whereas δ_z is a legitimate effect size for contrasts, it does not represent a natural generalization of Cohen's d measure.

Comparability. One of the consequences of the definition of the delta index computed with the z-method is that it may give surprising results when one is trying to translate an observed value to a planned research with a different design, an operation often done in power analysis. Consider, for instance, a case (cf. Cohen, 1988, p. 227) in which a researcher has observed a mean difference of 2.5 in a two-groups design, with $\boldsymbol{\mu_2} = \{4.5, 2\}$, such that $\delta_{z2} = 2.5/(\sigma\sqrt{2})$. She wishes to test the same difference in a three groups design in which she expects two groups to show the same mean of the first original group, and the third group as the second original group: that is, $\boldsymbol{\mu_3} = \{4.5, 4.5, 2\}$. She lays out the contrast weights $\boldsymbol{c} = \{1/2, 1/2, -1\}$. The expected contrast value (means comparison) is clearly equavalent in the two designs, namely $c\mu = 2.5$, but the expected effect size measure will be different. In the three groups design, she will obtain $\delta_{z3} = 2.5/(\sigma\sqrt{1.5})$. Thus, the comparison of the same mean values yields two different effect size outcomes. Obviously, the actual size of the contrast coefficients are immaterial here, because the index is standardized. The reason of the discrepancy in the example is that the standard deviation of the contrast decreases with increasing number of groups, even though the pooled standard deviation is the same across designs. Cohen (1988, pp. 276–278) discusses several cases in which he envisaged comparing designs with different number of groups sharing the same δ_0. In all these comparisons δ_z yields, somehow counterintuively, different effect sizes.

Inferential testability. The great advantage of the standardized contrast index δ_z is that it is easily associated with the inferential tests commonly used in testing the contrast hypothesis and to compute power parameters in power analysis. In fact, δ_z is strictly related to the t-test testing the contrast hypothesis. In particular:

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

which is the non-centrality parameter of the t-distribution used to compute the power function of the t-test associated with the contrast being examined (Liu, 2014; Steiger, 2004). This makes the standardization of the contrast a very handy scaling method when the main concern of the researcher is to compute power and power analysis is conducted with software that allows to specify the non-centrality parameter.

Furthermore, δ_z is strictly related also with two other effect size measures often used in power analysis and meta-analysis, the f and the η^2 (Cohen, 1988; Liu, 2013):

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

and

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

Scaled effect size measure

A different method of scaling the constrat effect size measure which guarantees better interpretability and comparability can be suggested. Let's $g = \frac{2}{\sum_i |c_i|}$, where $|c_i|$ indicates the absolute value of c_i, then

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

To be able to distinguish different effect size conceptualizations, we shall denote this measure of contrast effect size as δ_g and refer to it as computed with the g-method, for short. This method of scaling is equivalent to constraining the contrast weights such that $\sum |c_i| = 2$, suggested by a few authors (Kelley, 2018; Lai & Kelley, 2012). Thus, effect size indexes scaled with the g-method are in the same scale of indexes computed with such a constrain.

This scaling guarantees several advantages over the standardized effect size δ_z, although it is not devoid from nuisances.

Interpretation. The effect size computed with the g-method expresses the contrast value in terms of the pooled within-group standard deviation, but it constraints the contrast value to a comparable scale, the scale of the Cohen's d. Thus, it keeps the within-group standard deviation as the standardization scale, but makes the contrasts comparable across designs. In fact, in the special case of comparing two means, i.e. $\boldsymbol{c} = \{1, -1\}$, it yields exactly the same value of the Cohen's d.

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

Thus, it can be interpreted as the most used effect size measure in the literature. The particular choice of contrast weights does not matter, because the scaling compensates for the arbitrary parametrization of the contrast. In the two-group design, for instance, if one uses a contrast $\boldsymbol{c} = \{3, -3\}$, one obtains:

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

For larger designs with $k > 2$, the index keeps the expected meaning of Cohen's d, such that one can say that, for any number of groups k, a contrast with a given $\delta_{gk} = d$ shows the same effect size than two groups with a standardized mean difference of d.

In general terms, the g-method explicitly deals with the correct coding of the contrast weigths. Several authors have discussed standardized measures of contrast effect size, without explicitly defining how to transform an arbitrarily coded contrast into a properly coded one. Either in discussing examples (Kelley, 2007; Steiger, 2004), or in the implementation of software (Kelley, 2018, p. 30), it is often implicitily assumed that the contrast weights posses some characteristic which makes the standardized contrast interpretable as a Cohen's d. The g-method makes the transformation explicit, such that it can be employed with any arbitrary set of contrast weights.

Comparability. Consider, again, the case (cf. Cohen, 1988, p. 227) in which a researcher has observed a mean difference of 2.5 in a two-groups design, with $\boldsymbol{\mu_2} = \{4.5, 2\}$, such that $\delta_{02} = 2.5/\sigma$. In a three-group design he expects $\boldsymbol{\mu_3} = \{4.5, 4.5, 2\}$, and he evaluates the constrast $\boldsymbol{c} = \{1/2, 1/2, -1\}$. This is the case where the expected contrast value is equavalent in the two designs, namely $c\mu_3 = 2.5$. In virtue of the δ_g charateristics, in the three-group design the researcher will obtain

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

which is exactly what he obtains in the two groups design. Thus, the comparison of the same mean values yields identical effect size measures. One can also notice that the effect size quantity will be the same independently of the number of groups, k, provided the expected means are comparable. In particular, if one group has $\mu_1 = d$ and all other groups share the same mean $\mu_{2..k} = 0$, and the comparison is tested with the contrast $\boldsymbol{c} = \{(k-1), -1_2, -1_3, .., -1_k\}$, the contast value is $c\mu = (k-1) \cdot d$. The g-method yields, after simple algebra, $g = 1/(k-1)$, and thus $\delta_{gk} = \delta_{02}$, independently of the number of groups k.

It is easy to verify that in all the cases discussed by Cohen (1988, pp. 276–278) in which he envisaged comparing designs with different number of groups sharing the same δ_0, δ_g gives the same effect size quantity.

Inferential testability. One drawback of the scaled effect size measure is that it does not correspond directly to the inferential tests used to test hypotheses about the contrast or to compute the associated power function. Its correspondence, however, can be obtained by simple transformations. One needs to transform δ_g into a δ_z and exploit the latter index inferential testability properties. Thus, it is useful to relate the two indices:

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

It should be noted that the ratio z/g is usually very easy to compute. For an interaction contrast in a 2-by-2 design, $\boldsymbol{c} = \{1, -1 - 1, 1\}$, for instance, one can mentally verify that it is 4 divided by 4, that is 1. From equation (13) it follows that the expected value of the t-test associated with δ_g is:

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

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

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

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

Sample Estimation

As for the classical Cohen's d, estimating the population effect size using sample data may take different routes depending on the charateristics of the available data (Grissom & Kim, 2005). For δ-like measures, those routes differ in the way the within-group pooled standard deviation σ is estimated. It is important to clarify, however, that the way σ is estimated in the sample may change the numerical value of the effect size indices discussed here, but it does not alter their properties. In contrast analysis, the contrast is almost always estimated within the framework of the ANOVA, and the standard deviation σ is estimated as the square root of the mean square error of the ANOVA, namely (Howell, 2012, p. 204, p. 380; Kelley, 2007, Lai and Kelley (2012); Rosenthal et al., 2000, p. 41; Steiger, 2004):

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

which simplifies to $\sqrt{s^2}$ when groups have the same numerosity n and the same variance s^2. This choice of estimate of σ makes the effect size computed for the contrast equivalent to the Hedges's estimation of the standardized mean difference (Hedges & Olkin, 1985), often reffered to as *Hedges's g*. Other forms of estimation of σ are plausible, that can accomodate heteroskedasticity and different numerosity in the group, and reduce sampling bias (Grissom & Kim, 2005; Lai & Kelley, 2012).

Once s_p is available, and the population means are estimated with the sample means $\boldsymbol{m} = \{m_i\}$, for any given contrast $\boldsymbol{c} = \{c_i\}$, one can compute the scaled effect size indices. As regards estimating δ_z, one can lay out a contrast with arbitrarily scaled weights and compute:

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

Alternatively, one may first standardize the contrast weights and then compute $d = \sum_i (c_i m_i)/s_p$. Obviously, the latter method works because when the weigthts are standardized, $\sqrt{\sum c_i^2} = 1$, thus $\sum_i (c_i m_i)/s_p = d_z$. This also suggests that the z-method is handy for users of software that provides standardized contrasts weights by default, such as SPSS.

As regards estimating δ_g, one can lay out a contrast with arbitrarily scaled weights, and compute

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

or pick contrast weights that show $g = 1$. This can be achieved by constraining the weights such that $\sum |c_i| = 2$ (Lai & Kelley, 2012). Notice that there are several circumstances where choosing contrast weights with $g = 1$ is extremely easy. For instance, any contrast which 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[3] 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 matric. 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 η_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

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

The other two effect size indices defined by Rosenthal et al. (2000), namely $r_{effectsize}$ and r_{alert}, are not directly translatable into a d-like measure because they embed variance

[3]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)

which is not part of the within-group variance or the variance explained by the contrast, and thus they are not comparable with the proposed indices.

Power Analysis

A contrast hypothesis is usually tested with the t-test (Rosenthal et al., 2000; Steiger, 2004). Power analysis software usually provides a way to compute the power parameters (β, required n per cell, etc.) of a t-test based on the expected δ. The critical bit of information required for power calculation that is interesting here is the non-centrality parameter λ which affects the location of the t-distribution employed in computing power. For standard two independent samples t-test the parameter is (Cohen, 1988):

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

From the perspective of contrast analysis, λ_2 is a special case of a more general non-centrality parameter of the t-test with $k(n-1)$ degrees of freedom associated with a contrast $\boldsymbol{c}$ (Liu, 2014; Steiger, 2004):

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

which equates λ_2 when $\boldsymbol{c} = \{1, -1\}$ as in the two independent samples t-test. Because different scaling methods produce indices with different relations with δ_0, they require different transformations to obtain the correct non-centrality parameter.

When data are available for the computation of the non-centrality parameter one can simply estimate it as:

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

When only a scaled d is available, the non-centrality parameter needs to be scaled back to $\sqrt{n/\sum_i c_i^2} \cdot d_0$ accordingly to the used scaling method. When d is an estimation of δ_z,

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

When d is an estimation of δ_g,

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

In the last equation g is simply a constant that multiplies d_0 to obtain a scaled d_g. Thus, whatever scaling w one is using such that $\delta_w = \delta_0 \cdot w$, we have:

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

It may come as a surprise that different effect size quantities share the same non-centrality parameter and thus the same power function. However, this is always the case for effect size measures that refer to the same population parameter. In the linear model, for instance, the *B* and *beta* coefficients have very different scales, but they refer to the same population effect. Coherently, they lead to the same power function in power analysis.

A contrast hypothesis can also be tested with the F-test (Rosenthal et al., 2000). Power analysis software usually provides a way to compute the power parameters based on the F-test. The non-centrality parameter of the F-test is given by (Cohen, 1988, p. 481; Steiger, 2004):

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

thus, one simply transforms δ_z or δ_g into f and then compute the power of the F-test associated with 1 and $k(n-1)$ degrees of freedom.

Software usage

G*Power.

G*Power provides power functions for "generic t-test" which allows to input λ and *df* and returns the power. One can compute the non-centrality parameter as shown above and input it in the software. Unfortunately, the "generic t-test" function of the software does not allow to estimate the required N, an operation often useful for users. However, in G*Power one can compute all power parameters of a contrast using the F-test. Under "ANOVA: fixed effect, special, main effects and interactions" it is possible to specify k (number of groups), $df = k(n-1)$ and the effect size f. The correct f can be computed from d_z and d_g using equation (8) and (15).

R. To the best of our knowledge, R power functions commonly used in t-test power analysis do not allow to accomodate for contrasts with arbitrary weights. This is mainly due to the computation of df and λ that are tailored either to one-sample t-test, where $\lambda = \sqrt{N} \cdot \delta$ and $df = N-1$, or to two-samples t-test, where $\lambda = \sqrt{n} \cdot \delta$ and $df = N-2$

Thus, we have written a simple power function that computes the power parameters for contrasts based on estimated δ scaled in arbitrary ways, with shortcuts for the g-method and z-method. They are included in the `cpower` R package.

Confidence intervals

Confidence intervals for the estimated effect size indices can be computed using the *non-centrality interval estimation* method defined by Steiger and Fouladi (1997; se also Steiger, 2004). Their method entails to compute the confidence interval for the non-centrality parameter of the distribution associated with the effect size index, and then trasform the limits of that interval to the scale of the effect size.

When d is an estimation of δ_z, recall that the non-centrality parameter is equal to the observed t-test:

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

The lower ($\hat{\lambda}_l$) and the upper ($\hat{\lambda}_u$) limit of the $100(1-\alpha)\%$ confidence interval can be established by finding two noncentral distributions for which the value t represents the $100(1-\alpha/2)$-th and $100(\alpha/2)$-th percentile, respectively. This method yields the confidence interval in the scale of the non-centrality parameter:

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

When the interval is computed, one transforms the limits to scale them to the effect size scale (cf. Lai & Kelley, 2012):

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

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

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

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

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

and reduces to the Cohen's d confidence interval (Kelley, 2007), as expected. It is useful to reiterate that the confidence interval for d_z does not correspond to Cohen's d interval, simply because the two indices are generally different. It should be noted, however, that when the confidence interval is used to reject the null-hypothesis, both d_g and d_z lead to the same conclusion. In fact, the scaling of the noncentral parameter does not change the sign of the limits, thus if the interval around d_g contains zero, so does the interval around d_z, and viceversa. The R package `cpower` accompaning this contribution provides functions to compute confidence intervals for any contrast.

Examples

We now consider two examples to outline the methods discussed in this contribution in practical terms. We employ the R package `cpower` and make reference also to other software which can give the same results (See Appendix 1 for actual commands used in the examples).

In the first example we focus on computation of the effect size indices and their confidence intervals when data are available. Imagine a four groups design, with 30 participants in each group and each group representing an increasing level of a manipulated stimulus. The response to the stimulus has been recorded for each participant on a continuous scale. Data 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

The pooled standard deviation is of 7.23. We wish to estimate and quantify the linear and the quadratic trend of the four means. Thus, we lay out two sets of weights as described in Table 1.

As regards the linear trend, we obtain a statistically significant t-test, $t(116) = 5.93$, p.<.001, with a $d_g = 1.21$ and 95% confidence interval with limits [0.78,1.64]. We can then say that the linear trend shows a strong effect, equivalent to a Cohen's d of 1.21. More precisely, the average increment in mean response is 1.21 standard deviations across levels of stimulus. The standardized effect size is $d_z = 1.08$, with confidence limits [0.70,1.46]. As expected d_z is smaller than d_g, and its confidence interval slightly narrower.

As regards the quadratic trend, we obtain a statistically significant t-test, $t(116) = 2.65$, p.=0.01, with a $d_g = 0.48$ and 95% confidence interval with limits [0.12,0.85]. The quadratic trend shows a medium effect, equivalent to a Cohen's d of 0.48. The quadratic trend is much smaller than the linear trend, thus the increment of response across stimulus levels is more important than the curvature that the trend shows. The standardized effect size is $d_z = 0.48$, with confidence limits [0.12,0.85]. It is not surprising that for the quadratic trend we obtain $d_g = d_z$. In fact, in the quadratic $\sum |c_i^2| = 2 * \sqrt{\sum c_i^2}$ and thus $g = z$.

A second interesting example regards the computation of effect size indices in power analysis. The example is inspired by Wahlsten (1991) treatment of 2x2 designs. A researcher observes in a two-group design a factor A with means $\boldsymbol{m} = \{10, 7\}$ and $s_p = 4$. She wishes to run a 2 A x 2 B design where A is the same factor as in the two-group design, and B is a moderator. Let μ_{AB} be the expected mean for $A = \{1, 2\}$ and $B = \{1, 2\}$. She expects to replicate the two-group effect of A in condition $B = 1$, and absence of effect of A in condition $B = 2$. That is, $\{\mu_{11} = 10, \mu_{21} = 7, \mu_{12} = 7, \mu_{22} = 7\}$. Figure 1 represents the observed and the expected results.

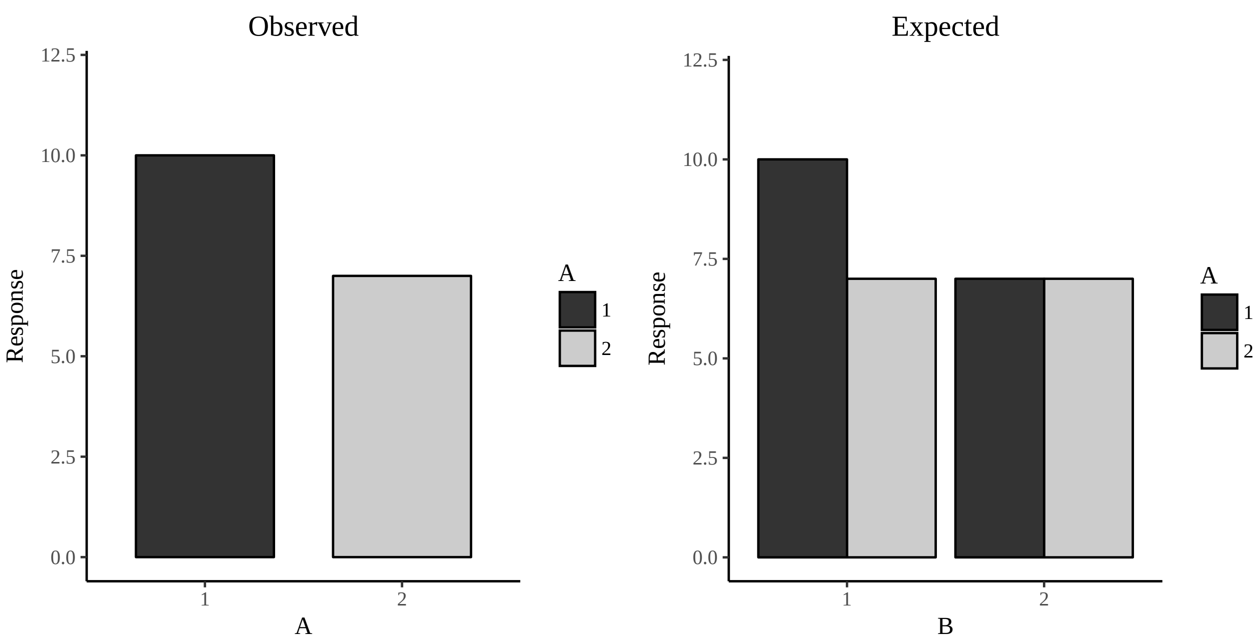

Figure 1. Observed and expected means in the example with two designs

The researcher wishes to estimate the expected δ_g for the interaction A X B, and its power parameters. For the 2-groups design, the observed effect size is:

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

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 δ_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 interpret 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.

Balancing advantages and disavantages, method-g δ_g index should be preferred over other scaling methods of the δ-like measures of effect size. It keeps the same scale of Cohen's d, and thus can be compared with published results. It also nicely generalizes the Cohen's d logic to multi-groups designs. {MORE good things here}

Nonetheless, the z-method scales can also be considered. This method of standardizing the contrasts weights has gather popularity and thus it can be often the case that published research provides, although implicitly, this form of effect indices. As long as the researcher uses this method consistently across designs, the method is useful for quick calculations of power parameters. Indeed, in meta-analysis both scaling methods can be used, provided that only one of them is used in the same analysis. When authors of published research have reported effect sizes with different scaling methods, the meta-analyst should convert them using the same scale. When published research includes effects reported as Cohen's d (and their variant), using the g-method is advisable, because it retains the 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(10,7,7,7)
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")
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")
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")
```

Example 2

```
library(cpower)
m2<-c(10,7)
c2<-(1,-1)
m4=c(10,7,7,7)
c4<-(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  ##
power.contrast.t(c4,dg,power = .80)
power.contrast.t(c2,dg*2,power = .80)

## transform dg in f  ##
f<-f.contr.d(linear,dg)
```

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.

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.