Behav Res (2018) 50:2057–2073
DOI 10.3758/s13428-017-0973-8

A methodological review of meta-analyses of the effectiveness of clinical psychology treatments

María Rubio-Aparicio[1] · **Fulgencio Marín-Martínez**[1] · **Julio Sánchez-Meca**[1] · **José Antonio López-López**[2]

Published online: 19 October 2017

Abstract This article presents a methodological review of 54 meta-analyses of the effectiveness of clinical psychological treatments, using standardized mean differences as the effect size index. We statistically analyzed the distribution of the number of studies of the meta-analyses, the distribution of the sample sizes in the studies of each meta-analysis, the distribution of the effect sizes in each of the meta-analyses, the distribution of the between-studies variance values, and the Pearson correlations between effect size and sample size in each meta-analysis. The results are presented as a function of the type of standardized mean difference: posttest standardized mean difference, standardized mean change from pretest to posttest, and standardized mean change difference between groups. These findings will help researchers design future Monte Carlo and theoretical studies on the performance of meta-analytic procedures, based on the manipulation of realistic model assumptions and parameters of the meta-analyses. Furthermore, the analysis of the distribution of the mean effect sizes through the meta-analyses provides a specific guide for the interpretation of the clinical significance of the different types of standardized mean differences within the field of the evaluation of clinical psychological interventions.

Keywords Meta-analysis · Standardized mean difference · Clinical significance · Monte Carlo studies

✉ Fulgencio Marín-Martínez
fulmarin@um.es

[1] Department of Basic Psychology & Methodology, Faculty of Psychology, University of Murcia, Murcia, Spain

[2] School of Social and Community Medicine, University of Bristol, Bristol, UK

Meta-analysis is a form of quantitative systematic review in which the results of a series of empirical studies on the same research topic are statistically summarized. When the individual studies report results in different scales (e.g., depression symptoms measured with different instruments), standardized effect size indices are often used to express the results across studies in a common metric. The standardized mean difference is one of the most used effect size indices in studies in which two or more groups are compared on a continuous outcome (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, Hedges, & Valentine, 2009).

The three main statistical objectives in a meta-analysis are to estimate the mean effect size through the primary studies, to assess the heterogeneity of the effect size estimates around the mean effect size, and to search for moderators that can explain part of the heterogeneity among the individual effect size estimates. In the behavioral, social, educational, and healthcare sciences, these moderators are the differential characteristics of the studies, such as the type of design, characteristics of the participant samples, or types of interventions (Hedges & Olkin, 1985; Rosenthal, 1991; Sánchez-Meca & Marín-Martínez, 2010).

There are two main statistical models with which to carry out a meta-analysis: the fixed-effect and the random-effects models. Under the *fixed-effect* model, it is assumed that all studies in the meta-analysis estimate a common population effect size, the only source of variability among the effect sizes being the sampling error due to the random selection of participants in each study (Konstantopoulos & Hedges, 2009). Conversely, in the *random-effects* model, it is assumed that each study in the meta-analysis estimates a different population effect size, and that studies are randomly selected from a population of studies, assuming that the corresponding population effect sizes are normally distributed. As a consequence, in the random-effects model, the effect sizes present two

sources of variability: between-studies and within-study variability. Furthermore, an additional assumption of the random-effects model is the independence between the sample sizes and the population effect sizes in the primary studies of the meta-analysis (Biggerstaff & Tweedie, 1997; Raudenbush, 2009).

Nowadays, there is a broad consensus that the random-effects model is more realistic than the fixed-effect model, due to the methodological and substantive differences that are typically found among the studies combined in a meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2010; Hedges & Vevea, 1998; National Research Council, 1992; Raudenbush, 1994, 2009). As a consequence, in this article we will focus on the methodological characteristics of random-effects model meta-analyses.

The empirical analysis of the methodological characteristics of real meta-analyses in a specific area of study is useful, as it helps to portrait the "typical" meta-analytic review that is conducted in a research field (e.g., number of studies, sample size distribution in the primary studies, and effect size distribution). Furthermore, a methodological review of meta-analyses allows assessing the degree of compliance with model assumptions, such as normal distribution of the effect sizes and independence between the sample sizes and effect sizes.

The aim of the present study was to explore the methodological characteristics of 54 meta-analyses published in high standard journals, which examined the effectiveness of clinical psychological interventions using standardized mean differences as the effect size index. This enabled us to provide a guide for the interpretation and characterization of the meta-analyses in the context of clinical psychology.

As in our study, Levine, Asada, and Carpenter (2009) explored the characteristics of 51 published meta-analyses on topics relevant to communication researchers (e.g., persuasion and interpersonal communication, language intensity effects, or viewing presidential debates). Interestingly, this study revealed a negative correlation between effect size and sample size for most of the meta-analyses reviewed, which may have been caused by publication bias.

Another review of meta-analyses was conducted by Engels, Schmid, Terrin, Olkin, and Lau (2000). These authors revised 125 published meta-analyses in the field of clinical medicine. They compared the performance of two effect size indices, the odds ratio and risk difference, usually applied in studies with binary outcomes. Both indices yielded the same conclusion when testing the statistical significance of the mean effect size within the same meta-analysis. However, risk differences led to greater heterogeneity than did odds ratios.

Schmidt, Oh, and Hayes (2009) selected 68 meta-analyses in which a fixed-effect model was assumed, and they reanalyzed the findings while applying the more realistic random-effects model. These meta-analyses focused on gender differences and the relations between personality and aggressive behavior. The fixed-effect confidence intervals around mean effect sizes showed an overstated and unrealistic precision, as compared to the wider random-effects confidence intervals.

Finally, Lipsey and Wilson (1993) reported an extensive review of meta-analyses of the efficacy of psychological and educational treatments. Some of the analyzed characteristics were the magnitude of the effects, the sample sizes of the primary studies, and the methodological quality of the meta-analyses. The main purpose of this study was to show the ability of meta-analysis to rigorously assess the degree of effectiveness of the treatments.

The present study focused on the methodological characteristics of meta-analyses of the effectiveness of treatments in the field of clinical psychology, with the standardized mean difference as the effect size index. Some of these methodological characteristics were the type of standardized mean difference (between groups or within groups), the distribution of the numbers of studies of the meta-analyses, the distribution of the sample sizes in the studies of each meta-analysis, the distribution of the effect sizes in each of the meta-analyses, the distribution of the between-studies variance values, and the Pearson correlations between the effect size and sample size in each meta-analysis.

With this methodological review of meta-analyses, we intend to offer a guide for the design of future research studies on the performance of meta-analytic procedures (e.g., Monte Carlo or theoretical studies), based on the manipulation of realistic assumptions and parameters in the meta-analyses. Furthermore, the analysis of the distribution of the average effect sizes through the meta-analyses will provide a guide for the interpretation of the clinical significance of the different types of standardized mean differences, in the field of the effectiveness of the clinical psychological treatments. In addition, our results will offer realistic estimates of effect size in this context, which is valuable information for researchers aiming to determine the optimal sample size when planning their investigations.

Types of standardized mean differences

If all of the studies included in the meta-analysis reported a continuous outcome in the same metric, raw mean differences could be used as the effect size index. However, this is seldom the case in the behavioral and social sciences, where different instruments to measure the same construct are usually considered across studies. This is why standardized mean differences are widely used in meta-analyses conducted in these fields.

Different types of standardized mean differences suit different study designs. In a two-group design (usually experimental vs. control) with a continuous outcome, the most usual

formula to estimate the population effect size is (as stated by Hedges & Olkin, 1985):

$$d = \left[1 - \frac{3}{4(n_1 + n_2) - 9}\right] \frac{\bar{y}_1 - \bar{y}_2}{\hat{S}}, \quad (1)$$

where d is an approximately unbiased estimator of the corresponding parameter, $\bar{y}_1$ and $\bar{y}_2$ are the means of the two groups for the outcome, n_1 and n_2 are the sample sizes, and $\hat{S}$ is an estimator of the pooled within-group standard deviation, which is given by

$$\hat{S} - \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}, \quad (2)$$

where S_1^2 and S_2^2 are the unbiased variances of the two groups.

Hedges and Olkin (1985) also derived the formula of the variance of the d index, $\hat{\sigma}_d^2$:

$$\hat{\sigma}_d^2 = \frac{n_1 + n_2}{n_1 \cdot n_2} + \frac{d^2}{2(n_1 + n_2)}. \quad (3)$$

In a repeated measures design, in which continuous pretest and posttest measures are registered for a sample of subjects (e.g., before and after the intervention), Becker (1988) proposed the standardized mean change, based on the difference between the pretest and posttest means divided by a standard deviation. Depending on the value of the estimated standard deviation in the denominator, there are two proposed d indices, which we will denote by d_{c1} and d_{c2}, respectively.

In a sample with n subjects, $\bar{y}_{pre}$ and $\bar{y}_{pos}$ denote the means in the pretest and posttest, respectively; d_{c1} is defined by Gibbons, Hedeker, and Davis (1993) as

$$d_{c1} = \left[1 - \frac{3}{4(n-1) - 1}\right] \frac{\bar{y}_{pre} - \bar{y}_{pos}}{S_c}, \quad (4)$$

where d_{c1} is an approximately unbiased estimator of the corresponding parameter; and S_c is the standard deviation of the change scores from pretest to posttest. The variance of d_{c1} has been given by Morris and DeShon (2002):

$$\hat{\sigma}_{d_{c1}}^2 = \left[1 - \frac{3}{4(n-1) - 1}\right]^2 \left(\frac{1}{n}\right)\left(\frac{n-1}{n-3}\right)\left(1 + n d_{c1}^2\right) - d_{c1}^2. \quad (5)$$

The d_{c2} index is given by Becker (1988), Morris (2000), and Morris and DeShon (2002):

$$d_{c2} = \left[1 - \frac{3}{4(n-1) - 1}\right] \frac{\bar{y}_{pre} - \bar{y}_{pos}}{S_{pre}}, \quad (6)$$

where S_{pre} is the standard deviation of the pretest scores, which is not influenced by the effects of the intervention. Morris (2000) derived the formula for estimating the variance of d_{c2}:

$$\hat{\sigma}_{d_{c2}}^2 = \left[1 - \frac{3}{4(n-1) - 1}\right]^2 \left(\frac{2(1-r)}{n}\right)\left(\frac{n-1}{n-3}\right)\left(1 + \frac{n d_{c2}^2}{2(1-r)}\right) - d_{c2}^2, \quad (7)$$

where r is the Pearson correlation between the pretest and posttest scores.

Note that, in studies with a two-independent-group design with continuous pretest and posttest measures, the most widely used effect size index is the standardized mean difference, d, as defined in Eq. 1, computed on the posttest scores. However, this index is only appropriate when assignment of the subjects to the groups is random and when equivalent pretest scores in both groups can be assumed. Furthermore, a disadvantage of computing the d index only on the posttest scores is that the valuable information of the pretest scores is ignored.

Becker (1988), Morris and DeShon (2002), and Morris (2008) proposed three effect size indices based on the difference between the standardized mean change in the experimental and control groups, which we will denominate d_{g1}, d_{g2}, and d_{g3}. These indices, unlike the standardized mean difference computed only on the posttest scores, take into account the information in both the pretest and posttest scores of the experimental and control groups.

The d_{g1} index is given by

$$d_{g1=} d_{c1,E} - d_{c1,C}, \quad (8)$$

where $d_{c1,E}$ and $d_{c1,C}$ are the standardized mean change, defined in Eq. 4, for the experimental and control groups, respectively. The variance of te d_{g1} can be estimated by

$$\hat{\sigma}_{d_{g1}}^2 = \hat{\sigma}_{d_{c1,E}}^2 + \hat{\sigma}_{d_{c1,C}}^2, \quad (9)$$

where $\hat{\sigma}_{d_{c1,E}}^2$ and $\hat{\sigma}_{d_{c1,C}}^2$ are the estimated variances of the d_{c1} indices computed by Eq. 5, applied to the experimental and control groups, respectively.

An alternative index to d_{g1} is d_{g2}, computed as the difference between the standardized mean change defined in Eq. 6 for the experimental and control groups:

$$d_{g2=} d_{c2,E} - d_{c2,C}. \quad (10)$$

The estimated variance of the d_{g2} index is given by

$$\hat{\sigma}_{d_{g2}}^2 = \hat{\sigma}_{d_{c2,E}}^2 + \hat{\sigma}_{d_{c2,C}}^2, \quad (11)$$

where $\hat{\sigma}^2_{d_{c2,E}}$ and $\hat{\sigma}^2_{d_{c2,C}}$ are the estimated variances of the d_{c2} indices computed by Eq. 7 for the experimental and control groups, respectively.

Assuming the homogeneity of the pretest standard deviations in the experimental and control groups, the d_{g3} index is given by

$$d_{g3} = \left[1-\frac{3}{4(n_E+n_c-2)-1}\right]\left[\frac{\left(\overline{y}_{pre,E}-\overline{y}_{pos,E}\right)-\left(\overline{y}_{pre,C}-\overline{y}_{pos,C}\right)}{\overline{S}_{pre}}\right], \quad (12)$$

where n_E and n_C are the sample sizes of the experimental and control groups, $\overline{y}_{pre,E}$ and $\overline{y}_{pos,E}$ are the means of the experimental group in the pretest and the posttest, $\overline{y}_{pre,C}$ and $\overline{y}_{pos,C}$ are the means of the control group in the pretest and the posttest, and $\overline{S}_{pre}$ is given by

$$\overline{S}_{pre} = \sqrt{\frac{(n_E-1)S^2_{pre,E}+(n_c-1)S^2_{pre,C}}{n_E+n_C-2}}, \quad (13)$$

where $S^2_{pre,E}$ and $S^2_{pre,C}$ are the variances of the experimental and control groups in the pretest.

Finally, the estimated variance of the d_{g3} index is given by

$$\hat{\sigma}^2_{d_{g3}} = 2\left[1-\frac{3}{4(n_E+n_C-2)-1}\right]^2(1-r)\left(\frac{n_E+n_C}{n_E n_C}\right)\left(\frac{n_E+n_C-2}{n_E+n_C-4}\right)\left[1+\frac{n_E n_C d^2_{g3}}{2(1-r)(n_E+n_C)}\right]-d^2_{g3} \quad (14)$$

where r is the mean of the Pearson correlations between the pretest and posttests scores in the experimental and control groups.

Methodology

Search procedure and selection criteria of the meta-analyses

The data for the present study were extracted from a sample of 50 published meta-analyses of the effectiveness of psychological treatments and interventions. The meta-analyses were obtained from journals with impact factors located in the first quartile of the 2011 Journal Citation Reports in the clinical psychology field (*Clinical Psychology Review*, *Psychological Medicine*, *Journal of Consulting and Clinical Psychology*, *Depression and Anxiety*, *Health Psychology*, *Neuropsychology*, *Behaviour Research and Therapy*, and *Journal of Substance Abuse Treatment*). The search was conducted in Google Scholar and limited to meta-analyses published between 2000 and 2012 with the keywords "meta-analysis" OR "systematic review" in the title.

First, reading the title and abstract of each reference allowed us to preselect meta-analyses of the effectiveness of different psychological programs, treatments, and interventions regarding psychological, educational, and psychosocial disorders. To be included in our study, meta-analyses had to comply with several selection criteria. First, we only included meta-analyses that used an effect size index from the *d* family: the posttest standardized mean difference (Eq. 1), standardized mean change (Eq. 4 or 6), and standardized mean change difference (Eq. 8, 10, or 12). Furthermore, the meta-analyses should report the individual effect sizes and sample sizes for the primary studies. To ensure that the selected meta-analyses had sufficient data to provide valid results, they had to include seven or more studies, with sample sizes of at least five subjects per group.

A total of 206 published meta-analyses were reviewed, of which 50 were finally included in the study. These included studies are marked with an asterisk in the References section. Some meta-analyses used two different effect sizes in the *d* family (Hesser, Weise, Westin, & Andersson, 2011; Nestoriuc, Rief, & Martin, 2008; Sockol, Epperson, & Barber, 2011; Virués-Ortega, 2010). In those cases, our decision was to consider them as independent meta-analyses. Thus, a total of 54 independent meta-analyses, or analysis units, were included in the present study. These meta-analyses summarized the results of 1,285 individual studies.

Data extraction

A database was created in SPSS, in which the effects sizes and sample sizes of the individual studies were coded for each meta-analysis. For meta-analyses including several outcomes, we selected the most relevant clinical outcome taking into account the principal aim of the meta-analysis. The type of design in which the computation of the effect size was based, and the type of *d* index were also recorded. Designs were classified as between-groups and within-groups, and type of *d* was coded as posttest standardized mean difference (*d* in Eq. 1), standardized mean change (d_{c1} or d_{c2}, in Eq. 4 or 6, respectively), and standardized mean change difference (d_{g1}, d_{g2}, or d_{g3}, in Eq. 8, 10, or 12, respectively). For each *d* value, its variance was estimated with Eq. 3, 5, 7, 9, 11, or 14, depending on the type of *d*.

The data from each meta-analysis were doubly coded by the first two authors of this article, with agreement percentages ranging between 94.44% and 100%. Inconsistencies between the coders were solved by consensus.

Meta-analytic calculations

Several computations were carried out using each meta-analytic database. The weighted average effect size was estimated using the following expression:

$$\overline{T} = \frac{\sum_i \hat{w}_i T_i}{\sum_i \hat{w}_i}, \tag{15}$$

where T_i refers to any d family effect size index, and $\hat{w}_i$ is the estimated weighting factor computed with the equation $\hat{w}_i = 1/\left(\hat{\sigma}_i^2 + \hat{\tau}_{DL}^2\right)$. The within-study variance of each individual study, $\hat{\sigma}_i^2$, was estimated using the formula corresponding to the type of d index (see Eqs. 3, 5, 7, 9, 11, and 14). The between-studies variance, $\hat{\tau}_{DL}^2$, was calculated using the procedure of DerSimonian and Laird (1986), the one most commonly used in practice. In this procedure, the between-studies variance estimator is derived from the moment method

$$\hat{\tau}_{DL}^2 = \frac{Q-(k-1)}{c}, \tag{16}$$

where k is the number of studies in the meta-analysis and Q is a statistic to test the heterogeneity of the effect sizes, described by Cochran (1954), and obtained by

$$Q = \sum_i \hat{w}_i^* \left(T_i - \overline{T}^*\right), \tag{17}$$

with $\hat{w}_i^*$ being the estimated weights assuming a fixed-effect model, $\hat{w}_i^* = 1/\hat{\sigma}_i^2$; $\overline{T}^*$ being the mean effect size also assuming a fixed-effect model—that is, applying Eq. 15, but using $\hat{w}_i^*$ as the weighting factor; and c being given by

$$c = \sum_i \hat{w}_i^* - \frac{\sum_i \left(\hat{w}_i^*\right)^2}{\sum_i \hat{w}_i^*}. \tag{18}$$

The mean effect size (Eq. 15) was always computed with the DL estimator. Restricted maximum likelihood (REML) and Paule and Mandel (PM) estimators of $\hat{\tau}^2$ were also applied, in order to know the distributions of the between-studies variances. Next we present formulas for these estimators.

The REML estimator is obtained iteratively from Sánchez-Meca and Marín-Martínez (2008) and Viechtbauer (2005):

$$\hat{\tau}_{REML}^2 = \frac{\sum_i \left(\hat{w}_i\right)^2 \left[\left(T_i - \overline{T}\right)^2 - \hat{\sigma}_i^2\right]}{\sum_i \left(\hat{w}_i\right)^2} + \frac{1}{\sum_i \hat{w}_i}, \tag{19}$$

where $\hat{w}_i$ is the estimated weighting factor, T_i refers to any d family effect size index, $\hat{\sigma}_i^2$ is the within-study variance of each individual study, and $\overline{T}$ is defined in Eq. 15. When $\hat{\tau}_{REML}^2 < 0$, it is truncated to zero.

The final estimator was also obtained through an iterative method, proposed by Paule and Mandel (1982). Applying this estimator, the between-studies variance is given by

$$\hat{\tau}_{PM}^2 = \sum_i \hat{w}_i \left(T_i - \overline{T}\right) - (k-1) \tag{20}$$

where $\hat{w}_i$ is the estimated weight, T_i is any of d family effect size, $\overline{T}$ is defined in Eq. 15, and k is the number of studies.

To test for true heterogeneity among the population effect sizes, we calculated the Q statistic, defined in Eq. 17, for each meta-analysis. Under the hypothesis of homogeneity among the effect sizes, the Q statistic follows a chi-square distribution with $k - 1$ degrees of freedom.

The Q statistic does not inform researchers of the extent of true heterogeneity, only of its statistical significance. Furthermore, the Q test has poor power to detect true heterogeneity among effect sizes when meta-analyses include a small number of studies ($k < 30$; Sánchez-Meca & Marín-Martínez, 1997). To overcome the shortcomings of the Q test, Higgins and Thompson (2002; Higgins, Thompson, Deeks, & Altman, 2003) proposed the I^2 index for assessing the magnitude of the heterogeneity exhibited by effect sizes. For each meta-analysis, the I^2 index was computed as

$$I^2 = \frac{Q-(k-1)}{Q} \times 100\% \tag{21}$$

The I^2 index was interpreted as the percentage of the total variability in a set of effect sizes due to true heterogeneity—that is, to between-studies variability. Indicatively, I^2 rates around 25%, 50%, and 75% can be interpreted as reflecting low, medium, and high heterogeneity, respectively (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006).

Data analysis

The statistical analyses were carried out in R. Specifically, the meta-analytic calculations were programmed with the metafor package (Viechtbauer, 2010), using the individual effect sizes and sample sizes coded for each meta-analysis as inputs. For repeated measures data, the correlation between the pre- and postassessments is required for computation of the variance of the d_{c2} (Eq. 7), d_{g2} (Eq. 11), and d_{g3} (Eq. 14) indices. Following Rosenthal (1991), the criterion was set at $r = .7$, as a representative value of the expected correlation in this context.

The normality assumption for the effect size distribution in each meta-analysis was assessed with the Shapiro–Wilk test for small samples and by computing the skewness and kurtosis of the distribution. Furthermore, the median, skewness, and kurtosis were also computed for the sample size

distribution in each meta-analysis. Descriptive analyses (minimum, maximum, mean, and quartiles) were carried out on the next indices across the meta-analyses: number of studies; mean effect size (Eq. 15); p value of the Shapiro–Wilk test; skewness and kurtosis of the d values; median, skewness, and kurtosis of the sample sizes distribution; Pearson correlation between effect sizes and sample sizes; and the p values of the heterogeneity Q statistic (Eq. 17), I^2 index (Eq. 21), and $\hat{\tau}^2$ index (Eqs. 16, 19, and 20). These analyses were performed separately for meta-analyses using the posttest standardized mean difference, the standardized mean change, and the standardized mean change difference.

The R code and a database of the 54 meta-analyses are available in the Open Science Framework (https://osf.io/yd52u/).

Results

Characteristics of the meta-analyses

A total of 54 meta-analyses were included in this study, of which 41 used the posttest standardized mean difference (between-groups design), 11 used the standardized mean change (within-groups design), and two used the standardized mean change difference (between-groups design). The database with the 54 meta-analyses is presented in the Appendix. The type of d family effect size index, the equation applied to estimate the variance of each individual effect size, and some meta-analytic calculations are recorded for each meta-analysis: number of studies; mean effect size; p value associated with the Q statistic; I^2; and $\hat{\tau}^2_{DL}$, $\hat{\tau}^2_{REML}$, and $\hat{\tau}^2_{PM}$ values. We performed these calculations using the values of the effect sizes and sample sizes from each meta-analysis.

The values of the d indices reported by the authors of the meta-analyses were computed as in Eqs. 1, 4, 6, 8, 10, and 12, or with some slight variations of these equations. Specifically, in some meta-analyses, the d_{c2} index (Eq. 6) was computed using pooled standard deviations from the pretest and posttest data, instead of the standard deviations in the pretest (i.e., the meta-analyses in Casement & Swanson, 2012; Driessen et al., 2010; Hansen, Höfling, Kröner-Borowik, Stangier, & Steil, 2013; and Williams, Hadjistavropoulos, & Sharpe, 2006). Also, in the meta-analysis of Aderka, Nickerson, Bøe, and Hofmann (2012), the d_{g3} index (Eqs. 12 and 13) was computed using the variances of the change scores instead of the variances in the pretest.

Some meta-analyses included more than one type of d index, and consequently, the 50 published meta-analyses were disaggregated into 54 independent meta-analyses. For

Table 1 Descriptive analyses of the meta-analytic calculations for posttest standardized mean difference

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
k	7	14	18	24.2	25	70
$\overline{d}$	0.068	0.249	0.409	0.472	0.695	1.075
p_norm	.000	.008	.138	.211	.312	.858
d_skewness	– 1.947	0.179	0.571	0.503	0.994	2.354
d_kurtosis	– 1.758	– 0.839	– 0.212	0.414	1.033	6.001
N_median	16	32	46.5	48.6	64	87.5
N_skewness	– 1.085	0.914	1.357	1.350	1.762	3.487
N_kurtosis	– 1.512	– 0.477	0.722	1.749	2.684	14.170
$r_{d,n}$	– .612	– .329	– .212	– .119	.059	.734
p_Q	.000	.000	.000	.095	.035	.981
I^2	0	37.71	59.86	54	74.83	93.61
$\hat{\tau}^2_{DL}$	.000	.055	.111	.159	.171	1.024
$\hat{\tau}^2_{REML}$	.000	.043	.108	.181	.179	.816
$\hat{\tau}^2_{PM}$	.000	.059	.129	.215	.352	.789

Min. = minimum; 1st Qu. = first quartile; 3rd Qu. = third quartile; Max. = maximum;

k = number of studies; $\overline{d}$ = average effect sizes applying DL to estimate the between-studies variance; p_norm = p value associated to the Shapiro–Wilk test; d_skewness = skewness of effect sizes; d_kurtosis = kurtosis of effect sizes; N_median = median of sample sizes ; N_skewness = skewness of sample sizes; N_kurtosis = kurtosis of sample sizes; $r_{d,n}$ = correlation between effect sizes and sample sizes; p_Q = p value associated to the heterogeneity Q statistic; I^2 = index to quantify the amount of heterogeneity (in %); $\hat{\tau}^2_{DL}$ = between-studies variance estimated using the DerSimonian and Laird (1986) method; $\hat{\tau}^2_{REML}$ = between-studies variance estimated using restricted maximum likelihood; $\hat{\tau}^2_{PM}$ = between-studies variance estimated using Paule and Mandel's (1982) method.

instance, as can be seen in the Appendix, the meta-analysis in Hesser et al. (2011) was disaggregated into two meta-analyses, since the standardized mean difference was used to compare the treatment and control groups at posttest, and the standardized mean change was used to evaluate the differences from pretest to posttest for some treatment groups.

Next, the distributions of the numbers of studies, effect sizes, sample sizes in the primary studies, correlations between effect sizes and sample sizes, and heterogeneity indices of the meta-analyses are presented as a function of the type of *d* index. Descriptive analyses of these distributions are shown for the meta-analyses using the posttest standardized mean difference (see Table 1), the standardized mean change (see Table 2), and the standardized mean change difference (see Table 3). Figure 1 shows the corresponding boxplots of the analyzed distributions for the meta-analyses using the posttest standardized mean differences (*d*) and the standardized mean changes (d_c), and Fig. 2 presents histograms of the mean effect sizes and between-studies variance distributions for the meta-analyses using the posttest standardized mean differences (*d*) and the standardized mean changes (d_c). Only two of the meta-analyses used the standardized mean change difference.

Number of studies

In the 41 meta-analyses that used the posttest standardized mean difference as the effect size index, the number of primary studies ranged from $k = 7$, the minimum number of studies for a meta-analysis to be included in this review, to $k = 70$. The first quartile, median, mean, and third quartile were 14, 18, 24.2, and 25 studies, respectively (see Table 1). These results reflect a clear positive skewness, or the predominance of meta-analyses with a small number of studies. Furthermore, as can be seen in Fig. 1, there were four outliers—namely, analyses of 45, 54, 61, and 70 studies—resulting in the mean, 24.2, being larger than the median, 18.

The distribution of the number of studies in the standardized mean change meta-analyses was more variable and more skewed than that in the posttest standardized mean difference meta-analyses (see Fig. 1). The first quartile, median, mean, and third quartile were 10, 13, 24.09, and 30 studies, respectively (see Table 2). These results evidenced a more pronounced positive skewness than in the case of the posttest standardized mean difference meta-analyses. Once again, most of the meta-analyses included only a small number of studies. The numbers of studies for the two meta-analyses using the standardized mean change difference were 9 and 19 (see Table 3).

Table 2 Descriptive analyses of the meta-analytic calculations for standardized mean change

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
k	8	10	13	24.09	30	70
$\overline{d}$	0.038	0.640	0.747	0.976	1.258	2.219
p_norm	.000	.001	.334	.320	.586	.830
d_skewness	– 1.179	– .0114	0.562	0.476	0.951	2.347
d_kurtosis	– 1.418	– 0.869	– 0.483	0.755	1.009	8.559
N_median	9	16	19.5	30.86	37.5	74
N_skewness	0.153	0.683	1.284	1.208	1.695	2.234
N_kurtosis	– 1.859	– 1.055	– 1.088	1.078	2.265	6.149
$r_{d,n}$	– .736	– .054	.045	.060	.318	.622
p_Q	.000	.000	.000	.002	.000	.013
I^2	44.99	64.86	72.67	72.74	81.61	93.46
$\hat{\tau}^2_{DL}$	.056	.099	.124	.185	.163	.512
$\hat{\tau}^2_{REML}$	.064	.105	.136	.211	.219	.588
$\hat{\tau}^2_{PM}$	.062	.088	.161	.299	.341	.588

Min. = minimum; 1st Qu. = first quartile; 3rd Qu. = third quartile; Max. = maximum;

k = number of studies; $\overline{d}$ = average effect sizes applying DL to estimate the between-studies variance; p_norm = p value associated to the Shapiro–Wilk test; d_skewness = skewness of effect sizes; d_kurtosis = kurtosis of effect sizes; N_median = median of sample sizes ; N_skewness = skewness of sample sizes; N_kurtosis = kurtosis of sample sizes; $r_{d,n}$ = correlation between effect sizes and sample sizes; p_Q = p-value associated to the heterogeneity Q statistic; I^2 = index to quantify the amount of heterogeneity (in %); $\hat{\tau}^2_{DL}$ = between-studies variance estimated using the DerSimonian and Laird (1986) method; $\hat{\tau}^2_{REML}$ = between-studies variance estimated using restricted maximum likelihood; $\hat{\tau}^2_{PM}$ = between-studies variance estimated using Paule and Mandel's (1982) method.

Effect size distributions

The mean effect size, the *p* value of the Shapiro–Wilk test for normality, and the skewness and kurtosis of the effect sizes were computed for each meta-analysis. To analyze the distribution of the mean effect sizes, these means were taken as absolute values. Note that the sign of a *d* index is arbitrary, since it depends on the order in which the means of the two groups in each primary study are subtracted. Thus, our interest was in the simple magnitudes of the mean effect sizes.

In the posttest standardized mean difference meta-analyses, the first quartile, median, mean, and third quartile of the mean effect size distribution were 0.249, 0.409, 0.472, and 0.695, respectively (see Table 1). These results are similar to the three values, 0.2, 0.5, and 0.8, that reflect low, medium, and high effect size magnitude, respectively, according to Cohen (1988).

The shape of the distribution of the posttest standardized mean differences in each meta-analysis was also examined. The Shapiro–Wilk test for normality was statistically significant for 39.02% of the meta-analyses, with .211 as the mean *p* value associated with this normality test. The skewness of distributions ranged from – 1.947 to 2.354, with 0.179 as the first quartile. Kurtosis ranged from – 1.758 to 6.001 (see Table 1). This means that the effect size distribution was positively skewed in most meta-analyses, with a statistically significant departure from normality in almost 40% of the meta-analyses.

In the meta-analyses using the standardized mean change, the three quartile, mean, and maximum values of the mean effect size distribution were larger than those for the posttest standardized mean difference meta-analyses (see Table 2 and Figs. 1 and 2). Specifically, the three quartiles were 0.640, 0.747, and 1.258, the mean was 0.976, and the maximum was 2.219, which was treated as an outlier. Note that these results remarkably exceed the 0.2, 0.5, and 0.8 values proposed by Cohen (1988).

Similar to posttest standardized mean differences, the shape of the standardized mean change distributions deviated from normality. This deviation was statistically significant in 36.36% of the meta-analyses, according to the Shapiro–Wilk test. The skewness and kurtosis distributions ranged from negative values in the first quartile to positive ones in the third quartile (see Table 2).

In the two meta-analyses using the standardized mean change difference, the mean effect sizes were 1.307 and 0.629, respectively (see Table 3). The skewness and kurtosis values were 0.383 and – 1.358 for the first meta-analysis, and – 0.514 and – 1.076 for the second meta-analysis. However, in both meta-analyses the Shapiro–Wilk test was not statistically significant, with the *p* values being .173 and .108, respectively.

Table 3 Meta-analytic calculations for standardized mean change difference

	Meta-Analysis 1	Meta-Analysis 2
k	9	19
$\overline{d}$	1.307	0.629
p_norm	.173	.108
d_skewness	0.383	– 0.514
d_kurtosis	– 1.358	– 1.076
N_median	28	38
N_skewness	1.026	1.745
N_kurtosis	0.006	2.296
$r_{d,n}$	.258	– .496
p_Q	.001	.000
I^2	69.49	68.45
$\hat{\tau}^2_{DL}$	.242	.109
$\hat{\tau}^2_{REML}$	.213	.083
$\hat{\tau}^2_{PM}$	.190	.066

Sample size distribution

We examined the sample size distributions through the *k* primary studies in each meta-analysis, by computing the median sample size, skewness, and kurtosis of the sample sizes. The distribution of these statistics was analyzed across the 41, 11, and two meta-analyses with different *d* effect size indices.

In the posttest standardized mean difference meta-analyses, the median sample size ranged from 16 to 87.5, with the mean being 48.6 (see Table 1). The first quartile of the skewness values was 0.914, which reflects a positive skewness of the sample size distributions in most meta-analyses (e.g., the primary studies predominantly had small sample sizes). The kurtosis values showed a large dispersion, ranging from – 1.512 to 14.170, with the first and third quartiles being – 0.477 and 2.684, respectively.

The sample sizes in the primary studies of the standardized mean change meta-analyses were lower than those in the posttest standardized mean difference meta-analyses (see Table 2 and Fig. 1). The median sample size ranged from 9 to 74 (an outlier), with a positively skewed distribution in which the three quartiles and the mean (16, 19.5, 37.5, and 30.86, respectively) were remarkably lower than those in the meta-analyses using the posttest standardized mean difference (32, 46.5, 64, and 48.6, respectively). The skewness values of the sample size distributions were all positive, ranging from 0.153 to 2.234, again suggesting the predominance of small sample sizes. The kurtosis values,

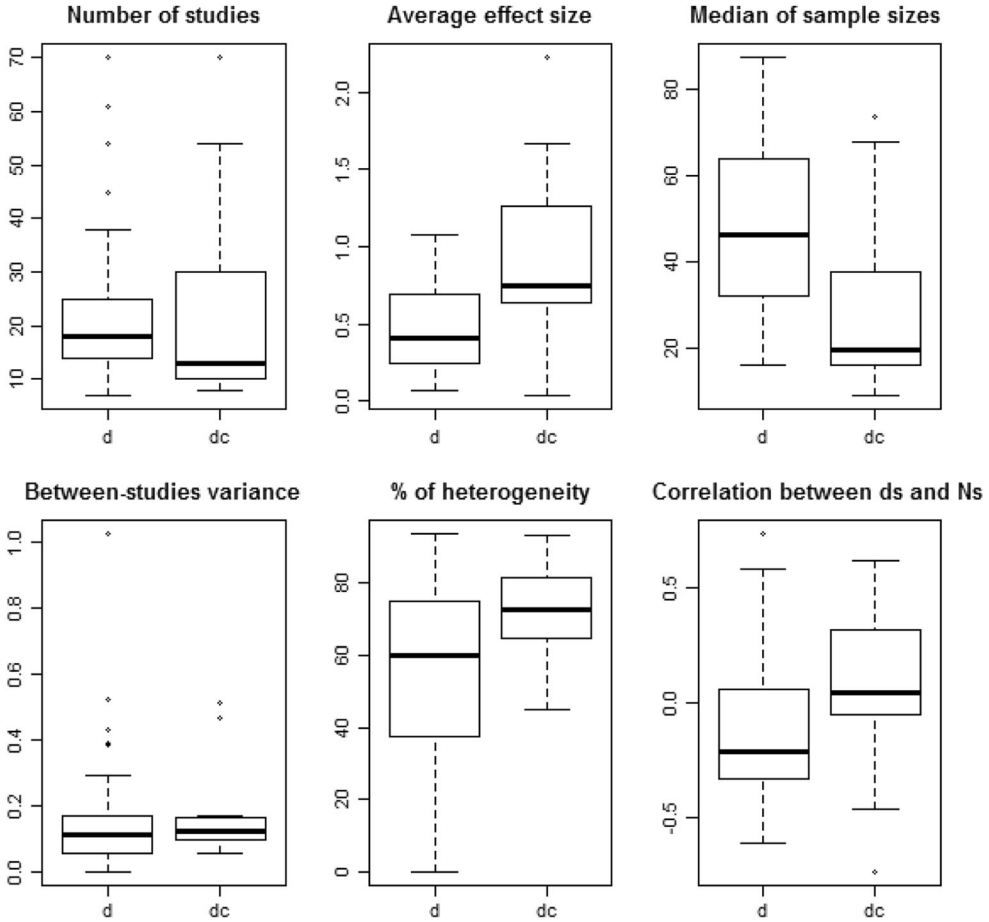

Fig. 1 Boxplots of some meta-analytic indices. Between-studies variance was estimated using the DerSimonian and Laird (1986) procedure. *d* = posttest standardized mean difference; d_c= standardized mean change

ranging from – 1.859 to 6.149, once again showed a large variability.

In the two standardized mean change difference meta-analyses, the medians of the sample sizes were 28 and 38, respectively (see Table 3). The skewnesses of the sample sizes were similar in the two meta-analyses, whereas the kurtosis values showed a higher discrepancy.

Correlation between effect sizes and sample sizes

Regarding meta-analyses using the posttest standardized mean difference, the correlations between effect sizes and sample sizes ranged from – .612 to .734. Most of the correlations (70.73%) were negative, with – .119 as the mean value (see Table 1). Out of the total of the correlations, 14.63% were statistically significant (three positive and three negative).

A wide range of correlations, from – .736 to .622, was also found in the standardized mean change meta-analyses (see Table 2 and Fig. 1). However, in this case most of the correlations (72.73%) were positive, and the mean of the correlations was also positive (.060). Out of the total number of correlations, 27.27% were statistically significant (two positive and one negative).

As is shown in Table 3, in the first meta-analysis the correlation between the standardized mean change differences and sample sizes was positive and not statistically significant (r = .258). In contrast, in the other meta-analysis the correlation was negative and statistically significant (r = – .496).

Heterogeneity

Three meta-analytic indices were used to study the heterogeneity of the effect sizes in the included meta-analyses: the Q statistic (Eq. 17), the I^2 index (Eq. 21), and the between-studies variance, $\hat{\tau}^2$, estimated using the DL, REML, and PM procedures (Eqs. 16, 19, and 20, respectively). Because the results for the three estimators of the between-studies variance were very similar, we will only describe the findings relative to the DL estimator, $\hat{\tau}^2_{DL}$.

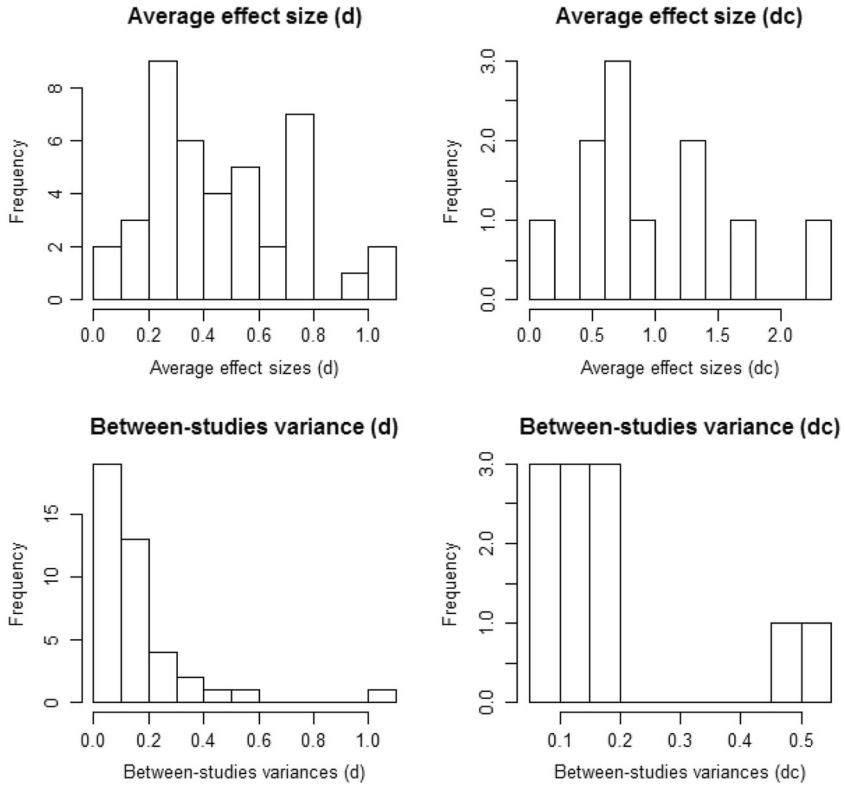

Fig. 2 Histograms of the distribution of mean effect sizes and between-studies variances for the posttest standardized mean difference (d) and standardized mean change (d_c). The between-studies variance was estimated using the DerSimonian and Laird (1986) procedure

In the meta-analyses using the posttest standardized mean difference, the third quartile of the distribution of p-values associated to the Q statistics was .035, below the .05 significance level (see Table 1). In particular, 75.6% of the Q tests were statistically significant at the .05 level (see the Appendix).

As is also shown in Table 1, the three quartiles of the I^2 distribution were 37.71%, 59.86%, and 74.83%, respectively. These results are relatively close to the three values, 25%, 50%, and 75%, proposed by Higgins and Thompson (2002) as reflecting low, medium, and high heterogeneity. Furthermore, 87.80% of the I^2 values were above 25%; that is to say, 36 of the 41 meta-analyses showed medium or high variability in the effect sizes (see the Appendix).

In the same vein, the three quartiles of the $\hat{\tau}^2_{DL}$ values were 0.055, 0.111, and 0.171, respectively, with four outliers in the distribution—namely, 1.025, 0.433, 0.522, and 0.384 (see Figs. 1 and 2).

The heterogeneity Q test was statistically significant in all of the standardized mean change meta-analyses (see Table 2). The I^2 values ranged from 44.99% to 93.46%, and the $\hat{\tau}^2_{DL}$ values ranged from 0.056 to 0.512. Thus, all the meta-analyses exhibited medium to large heterogeneity. As is shown in Fig. 1, the I^2 and $\hat{\tau}^2_{DL}$ values were generally larger in these meta-analyses, as compared to the posttest standardized mean difference meta-analyses.

The two meta-analyses using the standardized mean change difference showed statistically significant heterogeneity, with large I^2 and $\hat{\tau}^2_{DL}$ values, respectively above 68% and 0.10 (see Table 3).

Discussion

The aim of this article was to analyze the methodological characteristics of 54 meta-analyses of the effectiveness of psychological treatments in the clinical psychology area that used the standardized mean difference as the effect size index. These meta-analyses were extracted from the most high-impact journals in the field of clinical psychology, located in the first quartile of the rankings in the Journal Citation Reports.

The typical design in the primary studies evaluating the effectiveness of an intervention program was the pretest–posttest–control group design. Most meta-analyses, 41 out of 54, used the standardized mean difference computed from the posttest scores to compare experimental and control groups (Eq. 1). Eleven meta-analyses used the standardized mean change from pretest to posttest only in the treated groups (Eqs. 4 and 6), usually because some of the primary studies compared different treatments without including control groups. Finally, only two meta-analyses (Aderka et al., 2012,

and Virués-Ortega, 2010) used the standardized mean change difference proposed by Morris (2008) (Eq. 12), in which the gains from pretest to posttest are compared between the experimental and control groups.

The classic standardized mean difference computed from the posttest scores does not take into account the usual pre-differences between treatment and control groups, which also can occur in randomized studies. This poses a threat to the internal validity of the results. The standardized mean change from pretest to posttest in a treatment group can be affected by maturation, history or testing effects, which also represent a threat to the internal validity (Shadish, Cook, & Campbell, 2002). These limitations in both indices are partly overcome by using the standardized mean change difference between the experimental and control groups (Morris, 2008). Then, although in practice the standardized mean change difference is scarcely used (only in two out of our 54 meta-analyses), it should be considered in future meta-analyses.

The standardized mean change difference gives very similar results to those of the standardized mean difference computed from the change scores from pretest to posttest. This is especially true when the pretest scores are similar in the experimental and control groups. As a consequence, the methodological characteristics of our 41 meta-analyses using the posttest standardized mean difference can also serve to guide the design of future research on meta-analyses using the standardized mean change difference, as well as the correct interpretation of these meta-analyses.

In the global analysis of the 54 meta-analyses, many of them presented a relatively low number of studies (below 20) and a substantial heterogeneity of the effect sizes, with I^2 values being generally larger than 50%, and $\hat{\tau}^2$ values larger than 0.10. Thus, the performance of the meta-analytic statistical methods under these conditions should be a research topic of interest. It is widely known that the Q statistic for heterogeneity is underpowered in meta-analyses with a low number of studies (Sánchez-Meca & Marín-Martínez, 1997). However, in our review, 44 of the 54 meta-analyses showed statistically significant heterogeneity in the effect sizes ($p < .05$). This is because of the large I^2 and $\hat{\tau}^2$ values found in most of the meta-analyses, with only two meta-analyses showing $I^2 = 0\%$ and $\hat{\tau}^2 = 0$. These findings are in line with other studies supporting the random-effects model as a more realistic option than the fixed-effect model, on the basis that there is substantial variability in the effect sizes of meta-analyses (Hedges & Vevea, 1998; National Research Council, 1992; Raudenbush, 1994, 2009).

Cohen (1988) proposed a guide to interpret the magnitude of the standardized mean difference in the social sciences, where values around 0.2, 0.5, and 0.8, represent low, medium, and high effect size magnitudes, respectively. This guide should be adapted to the specific field of study, taking into account the typical distribution of effect sizes in the corresponding context (Ferguson, 2009; Hill, Bloom, Black, & Lipsey, 2008; Valentine & Cooper, 2003). In this vein, our study contributes to provide a tentative classification of the effect size magnitude of clinical psychology treatments, through the analysis of the distribution of mean effect sizes from our meta-analyses. Correct interpretation of the effect sizes in the empirical research makes it possible to determine the practical/clinical significance of the results, as a complement of the statistical significance (Kirk, 1996). Furthermore, the researcher can decide on the minimum effect size of interest to a priori determine the sample size of an empirical study, with the desired statistical power (Cohen, 1988).

The three quartiles of the mean effect size distribution were 0.249, 0.409, and 0.695 for the meta-analyses using the standardized mean difference computed from the posttest scores. These values, similar to those in Cohen (1988), could be interpreted as being of low, medium, and high magnitude, respectively, in the clinical psychology context. To be more specific: For example, a value of $d = 0.80$ could be interpreted as a high magnitude above the 75th percentile in the distribution of average effect sizes in the clinical psychology area. An important point is that this classification can only be applied to the posttest standardized mean differences and the standardized mean change differences, but not to the standardized mean changes from pretest to posttest.

Meta-analyses using the standardized mean change as an effect size index are more common than researchers would expect to find. This is because of the absence of control groups in the empirical research for ethical reasons, or when the studies are confined to comparing different active treatments without including a control group. In general terms, the values of the standardized mean change are larger than those of the posttest standardized mean difference. According to our review, the three quartiles of 0.64, 0.747, and 1.258 could be interpreted as a low, medium, and high magnitudes. This classification should be used instead of Cohen's (1988) proposal, for the interpretation of the standardized mean change values in the clinical psychological context.

The distribution of the effect sizes in the reviewed meta-analyses deviated from the normality assumption in the random-effects model. The skewness and kurtosis values ranged from negative to positive values of a remarkable magnitude, and the Shapiro–Wilk test for normality was statistically significant in almost 40% of the meta-analyses, in spite of the low number of studies in most of them, which reduced the statistical power of the test. These findings suggest the need to examine the robustness of the meta-analytic procedures to the violation of the normality assumption in the distribution of the effect sizes (see, e.g., Kontopantelis & Reeves, 2012), as well as the development of new, robust meta-analytic procedures.

The Pearson correlation between the effect sizes and sample sizes was statistically significant in ten out of the 54 meta-analyses, with five positive correlations and five negative. Once again, the low number of studies in numerous meta-analyses reduced the statistical power of the t test for the significance of a correlation, thus preventing the recognition of part of the true correlations. The distribution of the Pearson correlations in the meta-analyses, with values of a remarkable magnitude, could reflect publication selection bias, or possibly some other moderator confounded with sample size (e.g., implementation quality; Levine et al., 2009). As a consequence, the research about the performance of the meta-analytic procedures should consider scenarios with positive and negative correlation values between effect sizes and sample sizes, similar to those found in our study.

The present review of meta-analyses also provides the minimum and maximum, mean, and three quartiles of the distributions of the different components in a meta-analysis: number of studies; mean effect size; skewness and kurtosis of the effect size distribution; median, skewness and kurtosis of the sample size distribution; Pearson correlation between effect sizes and sample sizes; and the I^2 and $\hat{\tau}^2$ heterogeneity indices (see Tables 1, 2 and 3). These specific values are representative of the realistic conditions in a meta-analysis, which should be contemplated in the research about the performance of the meta-analytic procedures (see the next section for recommendations).

Limitations of the study

A requirement of the present review was to include only meta-analyses that reported individual effect sizes and sample sizes for the primary studies. That inclusion criterion might have led to the exclusion of meta-analyses with a large number of studies, due to journal space limitations. Nonetheless, our review included meta-analyses with numbers of studies ranging from seven to 70, which can be regarded as a wide range that realistically covers the size of most meta-analyses conducted in the social and behavioral sciences.

Only two meta-analyses out of the 54 in the review used the standardized mean change difference (Eq. 12), where the change scores from pretest to posttest between the experimental and control groups are compared. This index, although scarcely used in practice, overcomes some important limitations of the posttest standardized mean difference and the standardized mean change. As a consequence, it is suggested that future reviews include a larger number of meta-analyses using the standardized mean change difference.

This review was limited to meta-analyses of the effectiveness of clinical psychology treatments, using standardized mean differences as the effect index. Future reviews of meta-analyses in other research areas and with other effect size indices will shed light on the realistic meta-analytic conditions and typical distribution of the effect sizes in those disciplines.

Recommendation overview

Several recommendations can be made for researchers carrying out a meta-analysis, a Monte Carlo or theoretical study about meta-analytic methods, or a primary study. For studies with a pretest–posttest control group design, the best option is to compute the standardized mean change difference in each study, using Eq. 12. This index, although scarcely used in practice, has the advantage of controlling for pretest differences between the groups, as well as for maturation, history, or testing effects from pretest to posttest. Our article presents three indices of the standardized mean change difference: d_{g1}, d_{g2}, and d_{g3}, (Eqs. 8, 10, and 12, respectively), and the latter has been found to outperform the other indices in terms of bias, precision, and robustness to heterogeneity of variance (Morris, 2008).

The posttest standardized mean difference, d (Eq. 1), although widely applied in numerous meta-analyses, does not control for baseline differences between groups, which can also occur in randomized studies. However, in meta-analyses including studies with and without a pretest, the d index is the best option for all studies. This is because different standardized mean differences (e.g., posttest standardized mean differences, standardized mean changes from pretest to posttest, or standardized mean change differences) should not be combined in the same meta-analysis, since they are not directly comparable.

For studies with a pretest–posttest design without a control group, the usual approach is to compute a standardized mean change from pretest to posttest (d_{c1} and d_{c2} indices in Eqs. 4 and 6, respectively). These indices may be affected by maturation, history, or testing effects. However, in meta-analyses in which a sizeable number of studies do not include a control group, due to ethical reasons or only active treatments being compared, the d_c index could be computed in all studies. In this article we have presented two types of d_c indices that differ in the estimator of the standard deviation in the denominator of their formulas. The d_{c1} index (Eq. 4) uses the standard deviation of the change scores from pretest to posttest, whereas the d_{c2} index (Eq. 6) uses the standard deviation of the pretest scores. Most primary studies report the standard deviations of the pretest and posttest scores, whereas the standard deviation of the change scores is less frequently reported. Therefore, the computation of the d_{c2} index—based on the standard deviation of the pretest scores—will be more feasible in practice and will provide an estimation of the effect size more similar to those in the intergroup designs.

Monte Carlo and theoretical studies with a scope including meta-analytic methods should consider scenarios found in real meta-analyses. The results in Tables 1, 2 and 3 of this article can

inform the design of methodological studies in this context. For example, in a Monte Carlo study simulating data from meta-analyses using the posttest standardized mean difference or the standardized mean change difference, the number of studies, the sample size distribution in the primary studies, or the variance in the effect size distribution could be manipulated using the values in Table 1. For the number of studies, k, five values can be considered: 7, 14, 18, 25, and 70 (minimum, three quartiles, and maximum). Similarly, the sample size distribution could be manipulated with average values 16, 32, 46, 64, and 87 (minimum, three quartiles, and maximum), skewness of 1.357 (median), and kurtosis of .722 (median). Finally, the variance of the effect size distribution, $\hat{\tau}^2$, could be set to values of 0, 0.055, 0.111, 0.171, and 1.024 (minimum, the three quartiles, and maximum). These results may also be useful in a Bayesian framework, since they can define the construction of an empirical prior.

The distribution of average effect sizes throughout the reviewed meta-analyses can help researchers assess the practical significance (e.g., clinical significance) of an effect size in a primary empirical study or a meta-analysis in this context. For example, a value of $d = 0.20$ for the posttest standardized mean difference could be interpreted as a low magnitude, below the 25th percentile (0.249) in the distribution of the average effect sizes in clinical psychology (see Table 1). Furthermore, the benchmarks (minimum, Quartiles 1–3, and maximum) can help the researcher decide on the minimum effect size to determine a priori the sample size of an empirical study with the desired statistical power.

Conclusions

The results of this review of meta-analyses will allow proper interpretation of the magnitudes of the different types of standardized mean differences in the specific area of the evaluation of the effectiveness of the clinical psychological treatments. This is valuable information for interpreting the clinical significance of the results in both a primary research study and a meta-analysis, in terms of either the effect sizes of individual studies or the average effect sizes, both overall and by subgroups of studies, in a meta-analysis.

Future research on the performance of meta-analytic procedures should take into account the methodological characteristics of real meta-analyses in different areas of research. Particularly, in this work we have analyzed the number of studies, the sample size distribution in the studies, the effect size distribution, and the Pearson correlation between effect sizes and sample sizes of 54 real meta-analyses in the clinical psychology area. In this vein, Monte Carlo and theoretical studies could use the values reported in our study to simulate realistic scenarios.

Acknowledgements This article was supported by two grants from the Ministerio de Economía y Competitividad of the Spanish Government and by Fondo Europeo de Desarrollo Regional (Projects No. PSI2016-77676-P and PSI2015-71947-REDT).

Appendix

Table 4 Characteristics of the meta-analyses included in the systematic review

Meta-Analysis	d Index (Equation)	Formla for the Sampling Variance	k	$\bar{d}$	$\hat{\tau}^2_{DL}$	$\hat{\tau}^2_{REML}$	$\hat{\tau}^2_{PM}$	p	I^2
Abramowitz et al. (2001)	d (1)	Eq. 3	54	.250	.522	.625	.719	<.0001	81.3
Acarturk et al. (2009)	d (1)	Eq. 3	45	.740	.111	.104	.088	.0001	50.3
Aderka et al. (2012)	d_{g3} (12)	Eq. 13	19	.630	.109	.083	.066	<.0001	68.5
Bell & D'Zurilla (2009)	d (1)	Eq. 3	21	.694	.391	.560	.567	<.0001	83.6
Benish et al. (2008)	d (1)	Eq. 3	15	.187	.000	.000	.000	.9808	0
Burke et al. (2003)	d (1)	Eq. 3	13	.291	.022	.019	.020	.0824	37.7
Casement & Swanson (2012)	d_{c2} (6)	Eq. 7	13	.696	.090	.070	.062	<.0001	77.9
Cuijpers et al. (2009)	d (1)	Eq. 3	19	.307	.002	.012	.001	.4098	3.8
Cuijpers, Li , et al. (2010)	d (1)	Eq. 3	70	.195	.058	.056	.070	.0021	35.8
Cuijpers, Donker , et al. (2010)	d (1)	Eq. 3	24	.067	.098	.101	.099	.0093	45.1
Cuijpers et al. (2011)	d (1)	Eq. 3	15	.289	.000	.000	.000	.8662	0
Cuijpers et al. (2012)	d (1)	Eq. 3	18	.589	.008	.007	.008	.3501	8.8
Dixon et al. (2007)	d (1)	Eq. 3	20	.205	.008	.000	.014	.2493	16.4
Driessen et al. (2010)	d_{c2} (6)	Eq. 7	21	1.266	.152	.177	.173	<.0001	72.7
Ekers et al. (2008)	d (1)	Eq. 3	14	.072	.051	.038	.061	.1607	27.4
Gooding & Tarrier (2009)	d (1)	Eq. 3	18	.726	.163	.179	.164	<.0001	67.8
Hanrahan et al. (2013)	d (1)	Eq. 3	19	.928	.433	.689	.789	<.0001	81.8
Hansen et al. (2013)	d_{c2} (6)	Eq. 7	11	.563	.084	.106	.103	<.0001	83.1
Harris (2006)	d (1)	Eq. 3	14	.240	.081	.092	.109	.0004	65.3

Table 4 (continued)

Meta-Analysis	*d* Index (Equation)	Formla for the Sampling Variance	*k*	$\overline{d}$	$\hat{\tau}^2_{DL}$	$\hat{\tau}^2_{REML}$	$\hat{\tau}^2_{PM}$	*p*	I^2
Haug et al. (2012)	*d* (1)	Eq. 3	54	.799	.132	.130	.123	<.0001	71.8
Hausenblas et al. (2013)	d_{c1} (4)	Eq. 5	54	.038	.056	.064	.069	<.0001	62.9
Hesser et al. (2011) a	*d* (1)	Eq. 3	25	.600	.046	.043	.047	.0123	43.1
Hesser et al. (2011) b	d_{c2} (6)	Eq. 7	10	.584	.100	.104	.102	<.0001	90.2
Kalu et al. (2012)	*d* (1)	Eq. 3	7	.738	.272	.311	.389	.0266	58.0
Kleinstäuber et al. (2011)	*d* (1)	Eq. 3	18	.399	.142	.169	.192	<.0001	75.1
Lackner et al. (2004)	*d* (1)	Eq. 3	12	.766	.106	.109	.165	.0354	47.1
Lansbergen et al. (2007)	*d* (1)	Eq. 3	18	.220	.384	.636	.685	<.0001	86.1
Lissek et al. (2005)	*d* (1)	Eq. 3	22	.219	.130	.140	.162	.0002	59.9
Lundahl et al. (2006)	*d* (1)	Eq. 3	70	.463	.055	.058	.046	.0004	39.9
Malouff et al. (2007)	*d* (1)	Eq. 3	38	.546	.292	.538	.680	<.0001	83.6
Malouff et al. (2008)	*d* (1)	Eq. 3	15	.476	.125	.150	.163	<.0001	74.8
Nestoriuc et al. (2008) a	*d* (1)	Eq. 3	18	.298	.018	.015	.021	.2970	13.1
Nestoriuc et al. (2008) b	d_{c1} (4)	Eq. 5	70	.747	.124	.112	.161	<.0001	44.9
Oldham et al. (2012)	*d* (1)	Eq. 3	33	.378	.062	.065	.066	<.0001	65.0
Opriş et al. (2012)	*d* (1)	Eq. 3	23	.490	.255	.294	.352	<.0001	67.4
Pérez-Mañá et al. (2011)	*d* (1)	Eq. 3	21	.203	.081	.071	.059	<.0001	63.4
Prendergast et al. (2001)	*d* (1)	Eq. 3	11	.393	.157	.108	.083	<.0001	75.0
Richards & Richardson (2012)	*d* (1)	Eq. 3	33	.565	.141	.139	.129	<.0001	81.1
Roberts et al. (2007)	*d* (1)	Eq. 3	14	.363	.023	.011	.027	.1971	23.8
Rodenburg et al. (2009)	*d* (1)	Eq. 3	7	.560	.065	.068	.064	.1831	32.1
Rosa-Alcázar et al. (2008)	*d* (1)	Eq. 3	24	1.075	.173	.172	.378	.0002	57.9
Sánchez-Meca et al. (2010)	*d* (1)	Eq. 3	61	1.012	.261	.317	.363	<.0001	71.0
Shadish & Baldwin (2005)	*d* (1)	Eq. 3	30	.708	.160	.035	.431	.0014	49.3
Smit et al. (2012)	*d* (1)	Eq. 3	10	.331	1.024	.816	.789	<.0001	93.6
Sockol et al. (2011) a	*d* (1)	Eq. 3	14	.764	.171	.189	.194	<.0001	70.4
Sockol et al. (2011) b	d_{c1} (4)	Eq. 5	24	1.662	.467	.521	.510	<.0001	80.1
Spek et al. (2007)	*d* (1)	Eq. 3	11	.409	.077	.134	.145	<.0001	78.6
Sprenger et al. (2011)	*d* (1)	Eq. 3	10	.627	.114	.114	.147	.0174	55.2
Virués-Ortega (2010) a	d_{c2} (6)	Eq. 7	8	.984	.159	.179	.294	.0037	66.8
Virués-Ortega (2010) b	d_{g3} (12)	Eq. 13	9	1.307	.242	.213	.190	.0010	69.5
Westen & Morrison (2001)	d_{c2} (6)	Eq. 7	8	2.220	.512	.588	.589	<.0001	93.5
Williams et al. (2006)	d_{c2} (6)	Eq. 7	10	1.250	.117	.136	.073	.0131	56.9
Wittouck et al. (2011)	*d* (1)	Eq. 3	14	.163	.095	.105	.190	.0005	64.3
Young et al. (2007)	d_{c2} (6)	Eq. 7	36	.725	.167	.261	.387	<.0001	71.0

Note. a and b labels after a study indicate separate analyses of two sets of reported effect sizes. *d* = posttest standardized mean difference; d_{c1} = standardized mean change calculated using in the denominator the standard deviation of the pretest–posttest change scores; d_{c2} = standardized mean change calculated using in the denominator the standard deviation of the pretest scores; d_{g3} = standardized mean change difference calculated using in the denominator an average of the pretest standard deviations in the experimental and control groups; *k* = number of studies; $\overline{d}$ = mean effect size applying DL to estimate the between-studies variance; $\hat{\tau}^2_{DL}$ = between-studies variance estimated using the DerSimonian and Laird (1986) method; $\hat{\tau}^2_{REML}$ = between-studies variance estimated using restricted maximum likelihood; $\hat{\tau}^2_{PM}$ = between-studies variance estimated using Paule and Mandel's (1982) method; *p* = *p* value associated with the heterogeneity *Q* statistic; I^2 = index quantifying the amount of heterogeneity (%).

References

References preceded with an asterisk are those included in the methodological review.

*Abramowitz, J. S., Tolin, D. F., & Street, G. P. (2001). Paradoxical effects of thought suppression: A meta-analysis of controlled studies. *Clinical Psychology Review*, *21*, 683–703. doi:https://doi.org/10.1016/S0272-7358(00)00057-X

*Acarturk, C., Cuijpers, P., Van Straten, A., & De Graaf, R. (2009). Psychological treatment of social anxiety disorder: A meta-analysis. *Psychological Medicine*, *39*, 241–254. doi:https://doi.org/10.1017/S0033291708003590

*Aderka, I. M., Nickerson, A., Bøe, H. J., & Hofmann, S. G. (2012). Sudden gains during psychological treatments of anxiety and depression: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*, 93–101. doi:https://doi.org/10.1037/a0026455

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257–278. doi:https://doi.org/10.1111/j.2044-8317.1988.tb00901.x

*Bell, A. C., & D'Zurilla, T. J. (2009). Problem-solving therapy for depression: A meta-analysis. *Clinical Psychology Review*, *29*, 348–353. doi:https://doi.org/10.1016/j.cpr.2009.02.003

*Benish, S. G., Imel, Z. E., & Wampold, B. E. (2008). The relative efficacy of bona fide psychotherapies for treating post-traumatic stress disorder: A meta-analysis of direct comparisons. *Clinical Psychology Review*, *28*, 746–758. doi:https://doi.org/10.1016/j.cpr.2007.10.005

Biggerstaff, B. J., & Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, *16*, 753–768. doi:https://doi.org/10.1002/(SICI)1097-0258(19970415)16:7<753::AID-SIM494>3.3.CO;2-7

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: Wiley.

Borenstein, M., Hedges L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97–111. doi:https://doi.org/10.1002/jrsm.12

*Burke, B. L., Arkowitz, H., & Menchola, M. (2003). The efficacy of motivational interviewing: A meta-analysis of controlled clinical trials. *Journal of Consulting and Clinical Psychology*, *71*, 843–861. doi:https://doi.org/10.1037/0022-006X.71.5.843

*Casement, M. D., & Swanson, L. M. (2012). A meta-analysis of imagery rehearsal for post-trauma nightmares: Effects on nightmare frequency, sleep quality, and posttraumatic stress. *Clinical Psychology Review*, *32*, 566–574. doi:https://doi.org/10.1016/j.cpr.2012.06.002

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129. doi:https://doi.org/10.2307/3001666

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York: Russell Sage Foundation.

*Cuijpers, P., Clignet, F., van Meijel, B., van Straten, A., Li, J., & Andersson, G. (2011). Psychological treatment of depression in inpatients: A systematic review and meta-analysis. *Clinical Psychology Review*, *31*, 353–360. doi:https://doi.org/10.1016/j.cpr.2011.01.002

*Cuijpers, P., Donker, T., van Straten, A., Li, J., & Andersson, G. (2010). Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. *Psychological Medicine*, *40*, 1943–1957. doi:https://doi.org/10.1017/S0033291710000772

*Cuijpers, P., Driessen, E., Hollon, S. D., van Oppen, P., Barth, J., & Andersson, G. (2012). The efficacy of non-directive supportive therapy for adult depression: A meta-analysis. *Clinical Psychology Review*, *32*, 280–291. doi:https://doi.org/10.1016/j.cpr.2012.01.003

*Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review*, *30*, 768–778. doi:https://doi.org/10.1016/j.cpr.2010.06.001

*Cuijpers, P., van Straten, A., Warmerdam, L., & Andersson, G. (2009). Psychotherapy versus the combination of psychotherapy and pharmacotherapy in the treatment of depression: A meta-analysis. *Depression and Anxiety*, *26*, 279–288. doi:https://doi.org/10.1002/da.20519

DerSimonian, R., & Laird, N. (1986). Meta-analysis of clinical trials. *Controlled Clinical Trials*, *7*, 177–188. doi:https://doi.org/10.1016/0197-2456(86)90046-2

*Dixon, K. E., Keefe, F. J., Scipio, C. D., Perri, L. M., & Abernethy, A. P. (2007). Psychological interventions for arthritis pain management in adults: A meta-analysis. *Health Psychology*, *26*, 241–250. doi:https://doi.org/10.1037/0278-6133.26.3.241

*Driessen, E., Cuijpers, P., de Maat, S. C., Abbass, A. A., de Jonghe, F., & Dekker, J. J. (2010). The efficacy of short-term psychodynamic psychotherapy for depression: A meta-analysis. *Clinical Psychology Review*, *30*, 25–36. doi:https://doi.org/10.1016/j.cpr.2009.08.010

*Ekers, D., Richards, D., & Gilbody, S. (2008). A meta-analysis of randomized trials of behavioural treatment of depression. *Psychological Medicine*, *38*, 611–623. doi:https://doi.org/10.1017/S0033291707001614

Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Statistics in Medicine*, *19*, 1707–1728. doi:https://doi.org/10.1002/1097 0258(20000715)19:13<1707::AID-SIM491>3.0.CO;2-P

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*, 532–538. doi:https://doi.org/10.1037/a0015808

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational and Behavioral Statistics*, *18*, 271–279. doi:https://doi.org/10.2307/1165136

*Gooding, P., & Tarrier, N. (2009). A systematic review and meta-analysis of cognitive–behavioural interventions to reduce problem gambling: Hedging our bets? *Behaviour Research and Therapy*, *47*, 592–607. doi:https://doi.org/10.1016/j.brat.2009.04.002

*Hanrahan, F., Field, A. P., Jones, F. W., & Davey, G. C. (2013). A meta-analysis of cognitive therapy for worry in generalized anxiety disorder. *Clinical Psychology Review*, *33*, 120–132. doi:https://doi.org/10.1016/j.cpr.2012.10.008

*Hansen, K., Höfling, V., Kröner-Borowik, T., Stangier, U., & Steil, R. (2013). Efficacy of psychological interventions aiming to reduce chronic nightmares: A meta-analysis. *Clinical Psychology Review*, *33*, 146–155. doi:https://doi.org/10.1016/j.cpr.2012.10.012

*Harris, A. H. (2006). Does expressive writing reduce health care utilization? A meta-analysis of randomized trials. *Journal of Consulting and Clinical Psychology*, *74*, 243–252. doi:https://doi.org/10.1037/0022-006X.74.2.243

*Haug, T., Nordgreen, T., Öst, L. G., & Havik, O. E. (2012). Self-help treatment of anxiety disorders: A meta-analysis and meta-regression of effects and potential moderators. *Clinical Psychology Review*, *32*, 425–445. doi:https://doi.org/10.1016/j.cpr.2012.04.002

*Hausenblas, H. A., Campbell, A., Menzel, J. E., Doughty, J., Levine, M., & Thompson, J. K. (2013). Media effects of experimental presentation of the ideal physique on eating disorder symptoms: A meta-analysis of laboratory studies. *Clinical Psychology Review*, *33*, 168–181. doi:https://doi.org/10.1016/j.cpr.2012.10.011

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504. doi:https://doi.org/10.1037/1082-989X.3.4.486

*Hesser, H., Weise, C., Westin, V. Z., & Andersson, G. (2011). A systematic review and meta-analysis of randomized controlled trials of cognitive–behavioral therapy for tinnitus distress. *Clinical Psychology Review*, *31*, 545–553. doi:https://doi.org/10.1016/j.cpr.2010.12.006

Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*, 1539–1558. doi:https://doi.org/10.1002/sim.1186

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*, 557–560. doi:https://doi.org/10.1136/bmj.327.7414.557

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177. doi:https://doi.org/10.1111/j.1750-8606.2008.00061.x

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2

index? *Psychological Methods*, *11*, 193–206. doi:https://doi.org/10.1037/1082-989X.11.2.193

*Kalu, U. G., Sexton, C. E., Loo, C. K., & Ebmeier, K. P. (2012). Transcranial direct current stimulation in the treatment of major depression: A meta-analysis. *Psychological Medicine*, *42*, 1791–1800. doi:https://doi.org/10.1017/S0033291711003059

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759. doi:https://doi.org/10.1177/0013164496056005002

*Kleinstäuber, M., Witthöft, M., & Hiller, W. (2011). Efficacy of short-term psychotherapy for multiple medically unexplained physical symptoms: A meta-analysis. *Clinical Psychology Review*, *31*, 146–160. doi:https://doi.org/10.1016/j.cpr.2010.09.001

Konstantopoulos, S., & Hedges, L. V. (2009). Analyzing effect sizes: Fixed-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 279–293). New York, NY: Russell Sage Foundation.

Kontopantelis, E., & Reeves, D. (2012). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical Methods in Medical Research*, *21*, 409–426. doi:https://doi.org/10.1177/0962280210392008

*Lackner, J. M., Mesmer, C., Morley, S., Dowzer, C., & Hamilton, S. (2004). Psychological treatments for irritable bowel syndrome: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *72*, 1100–1113. doi:https://doi.org/10.1037/0022-006X.72.6.1100

*Lansbergen, M. M., Kenemans, J. L., & van Engeland, H. (2007). Stroop interference and attention-deficit/hyperactivity disorder: A review and meta-analysis. *Neuropsychology*, *21*, 251–262. doi:https://doi.org/10.1037/0894-4105.21.2.251

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample sizes and effect sizes are negatively correlated in meta-analyses: Evidence and implications of a publication bias against nonsignificant findings. *Communication Monographs*, *76*, 286–302. doi:https://doi.org/10.1080/03637750903074685

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*, 1181–1209. doi:https://doi.org/10.1037/0003-066X.48.12.1181

*Lissek, S., Powers, A. S., McClure, E. B., Phelps, E. A., Woldehawariat, G., Grillon, C., & Pine, D. S. (2005). Classical fear conditioning in the anxiety disorders: A meta-analysis. *Behaviour Research and Therapy*, *43*, 1391–1424. doi:https://doi.org/10.1016/j.brat.2004.10.007

*Lundahl, B., Risser, H. J., & Lovejoy, M. C. (2006). A meta-analysis of parent training: Moderators and follow-up effects. *Clinical Psychology Review*, *26*, 86–104. doi:https://doi.org/10.1016/j.cpr.2005.07.004

*Malouff, J. M., Thorsteinsson, E. B., Rooke, S. E., Bhullar, N., & Schutte, N. S. (2008). Efficacy of cognitive behavioral therapy for chronic fatigue syndrome: A meta-analysis. *Clinical Psychology Review*, *28*, 736–745. doi:https://doi.org/10.1016/j.cpr.2007.10.004

*Malouff, J. M., Thorsteinsson, E. B., & Schutte, N. S. (2007). The efficacy of problem solving therapy in reducing mental and physical health problems: A meta-analysis. *Clinical Psychology Review*, *27*, 46–57. doi:https://doi.org/10.1016/j.cpr.2005.12.005

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, *53*, 17–29. doi:https://doi.org/10.1348/000711000159150

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*, 364–386. doi:https://doi.org/10.1177/1094428106291059

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105–125. doi:https://doi.org/10.1037/1082-989X.7.1.105

National Research Council (1992). *Combining information: Statistical issues and opportunities for research.* Washington, DC: National Academy Press.

*Nestoriuc, Y., Rief, W., & Martin, A. (2008). Meta-analysis of biofeedback for tension-type headache: Efficacy, specificity, and treatment moderators. *Journal of Consulting and Clinical Psychology*, *76*, 379–396. doi:https://doi.org/10.1037/0022-006X.76.3.379

*Oldham, M., Kellett, S., Miles, E., & Sheeran, P. (2012). Interventions to increase attendance at psychotherapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology*, *80*, 928–939. doi:https://doi.org/10.1037/a0029630

*Opriș, D., Pintea, S., García-Palacios, A., Botella, C., Szamosközi, Ș., & David, D. (2012). Virtual reality exposure therapy in anxiety disorders: A quantitative meta-analysis. *Depression and Anxiety*, *29*, 85–93. doi:https://doi.org/10.1002/da.20910

Paule, R. C., & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, *87*, 377–385.

*Pérez-Mañá, C., Castells, X., Vidal, X., Casas, M., & Capellà, D. (2011). Efficacy of indirect dopamine agonists for psychostimulant dependence: A systematic review and meta-analysis of randomized controlled trials. *Journal of Substance Abuse Treatment*, *40*, 109–122. doi:https://doi.org/10.1016/j.jsat.2010.08.012

*Prendergast, M. L., Urada, D., & Podus, D. (2001). Meta-analysis of HIV risk-reduction interventions within drug abuse treatment programs. *Journal of Consulting and Clinical Psychology*, *69*, 389–405. doi:https://doi.org/10.1037/0022-006X.69.3.389

Raudenbush, S. W. (1994). Random effects models. In H. Cooper, & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 295–315). New York: Russell Sage Foundation.

*Richards, D., & Richardson, T. (2012). Computer-based psychological treatments for depression: A systematic review and meta-analysis. *Clinical Psychology Review*, *32*, 329–342. doi:https://doi.org/10.1016/j.cpr.2012.02.004

*Roberts, M. E., Tchanturia, K., Stahl, D., Southgate, L., & Treasure, J. (2007). A systematic review and meta-analysis of set-shifting ability in eating disorders. *Psychological Medicine*, *37*, 1075–1084. doi:https://doi.org/10.1017/S0033291707009877

*Rodenburg, R., Benjamin, A., de Roos, C., Meijer, A. M., & Stams, G. J. (2009). Efficacy of EMDR in children: A meta-analysis. *Clinical Psychology Review*, *29*, 599–606. doi:https://doi.org/10.1016/j.cpr.2009.06.008

*Rosa-Alcázar, A. I., Sánchez-Meca, J., Gómez-Conesa, A., & Marín-Martínez, F. (2008). Psychological treatment of obsessive–compulsive disorder: A meta-analysis. *Clinical Psychology Review*, *28*, 1310–1325. doi:https://doi.org/10.1016/j.cpr.2008.07.001

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.) Newbury Park: Sage.

Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. *Quality and Quantity*, *31*, 385–399. doi:https://doi.org/10.1023/A:1004298118485

Sánchez-Meca J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods*, *13*, 31–48. doi:https://doi.org/10.1037/1082-989X.13.1.31

Sánchez-Meca, J., & Marín-Martínez, F. (2010). Meta-analysis. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., Vol. 7, pp. 274–282). Oxford: Elsevier.

*Sánchez-Meca, J., Rosa-Alcázar, A. I., Marín-Martínez, F., & Gómez-Conesa, A. (2010). Psychological treatment of panic disorder with or without agoraphobia: A meta-analysis. *Clinical Psychology Review*, *30*, 37–50. doi:https://doi.org/10.1016/j.cpr.2009.08.011

Schmidt, F. L., Oh, I. S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*, 97–128. doi:https://doi.org/10.1348/000711007X255327

*Shadish, W. R., & Baldwin, S. A. (2005). Effects of behavioral marital therapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology*, *73*, 6–14. doi:https://doi.org/10.1037/0022-006X.73.1.6

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

*Smit, Y., Huibers, M. J., Ioannidis, J. P., van Dyck, R., van Tilburg, W., & Arntz, A. (2012). The effectiveness of long-term psychoanalytic psychotherapy—A meta-analysis of randomized controlled trials. *Clinical Psychology Review*, *32*, 81–92. doi:https://doi.org/10.1016/j.cpr.2011.11.003

*Sockol, L. E., Epperson, C. N., & Barber, J. P. (2011). A meta-analysis of treatments for perinatal depression. *Clinical Psychology Review*, *31*, 839–849. doi:https://doi.org/10.1016/j.cpr.2011.03.009

*Spek, V., Cuijpers, P., Nyklícek, I., Riper, H., Keyzer, J., & Pop, V. (2007). Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: A meta-analysis. *Psychological Medicine*, *37*, 319–328. doi:https://doi.org/10.1017/S0033291706008944

*Sprenger, L., Gerhards, F., & Goldbeck, L. (2011). Effects of psychological treatment on recurrent abdominal pain in children—A meta-analysis. *Clinical Psychology Review*, *31*, 1192–1197. doi:https://doi.org/10.1016/j.cpr.2011.07.010

Valentine, J. C., & Cooper, H. M. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Washington, DC: What Works Clearinhouse.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*, 261–293. doi:https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48.

*Virués-Ortega, J. (2010). Applied behavior analytic intervention for autism in early childhood: Meta-analysis, meta-regression and dose–response meta-analysis of multiple outcomes. *Clinical Psychology Review*, *30*, 387–399. doi:https://doi.org/10.1016/j.cpr.2010.01.008

*Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *69*, 875. doi:https://doi.org/10.1037/0022-006X.69.6.875

*Williams, J., Hadjistavropoulos, T., & Sharpe, D. (2006). A meta-analysis of psychological and pharmacological treatments for body dysmorphic disorder. *Behaviour Research and Therapy*, *44*, 99–111. doi:https://doi.org/10.1016/j.brat.2004.12.006

*Wittouck, C., Van Autreve, S., De Jaegere, E., Portzky, G., & van Heeringen, K. (2011). The prevention and treatment of complicated grief: A meta-analysis. *Clinical Psychology Review*, *31*, 69–78. doi:https://doi.org/10.1016/j.cpr.2010.09.005

*Young, K. M., Northern, J. J., Lister, K. M., Drummond, J. A., & O'Brien, W. H. (2007). A meta-analysis of family-behavioral weight-loss treatments for children. *Clinical Psychology Review*, *27*, 240–249. doi:https://doi.org/10.1016/j.cpr.2006.08.003