Psychotherapy Research

ISSN: 1050-3307 (Print) 1468-4381 (Online) Journal homepage: http://www.tandfonline.com/loi/tpsr20

Effect size calculation in meta-analyses of psychotherapy outcome research

William T. Hoyt & A. C. Del Re

To cite this article: William T. Hoyt & A. C. Del Re (2017): Effect size calculation in meta-analyses of psychotherapy outcome research, Psychotherapy Research, DOI: 10.1080/10503307.2017.1405171

To link to this article: https://doi.org/10.1080/10503307.2017.1405171

Psychotherapy Research, 2017
https://doi.org/10.1080/10503307.2017.1405171

Considerations of How to Conduct Meta-Analyses in Psychological Interventions

Effect size calculation in meta-analyses of psychotherapy outcome research

WILLIAM T. HOYT[1] & A. C. DEL RE[2]

[1]*University of Wisconsin-Madison, Madison, WI, USA &* [2]*Palo Alto Veterans Administration Medical Center, Palo Alto, CA, USA*

(Received 17 July 2017; revised 30 September 2017; accepted 20 October 2017)

Abstract
Meta-analysis of psychotherapy intervention research normally examines differences between treatment groups and some form of comparison group (e.g., wait list control; alternative treatment group). The effect of treatment is normally quantified as a standardized mean difference (SMD). We describe procedures for computing unbiased estimates of the population SMD from sample data (e.g., group *M*s and SDs), and provide guidance about a number of complications that may arise related to effect size computation. These complications include (a) incomplete data in research reports; (b) use of baseline data in computing SMDs and estimating the population standard deviation (σ); (c) combining effect size data from studies using different research designs; and (d) appropriate techniques for analysis of data from studies providing multiple estimates of the effect of interest (i.e., *dependent* effect sizes).

Keywords: meta-analysis; effect size computation; dependent effect sizes

Clinical or Methodological Significance of this article: Meta-analysis is a set of techniques for producing valid summaries of existing research. The initial computational step for meta-analyses of research on intervention outcomes involves computing an *effect size* quantifying the change attributable to the intervention. We discuss common issues in the computation of effect sizes and provide recommended procedures to address them.

Meta-analysis encompasses a family of techniques for synthesis of quantitative research findings. Based on a comprehensive literature search, the meta-analyst retrieves the available studies that address the research question of interest, then extracts an *effect size* from each study summarizing that study's findings regarding the strength of association between the independent (or predictor) variable and the dependent (or outcome) variable (Cooper & Hedges, 2009). This article reviews best practices for effect size calculation for researchers conducting meta-analyses of psychotherapy outcome research, focusing on challenges, such as (a) incomplete reporting of quantitative data; (b) effect size computation with or without baseline data; (c) studies with multiple comparison groups; and (d) studies with multiple outcome measures.

We assume that the outcome is a continuous (numeric) score, so that the natural measure of effect size is some type of standardized mean difference (SMD), which quantifies the average difference between two groups (labeled *treatment* and *comparison*) in standardized units. This produces a standardized effect size which is readily interpretable (e.g., SMD = 1 reflects a difference of 1 SD between treatment and comparison groups), allowing for meaningful comparisons between studies using different outcome measures.

When a study does not include data on the type of comparison group (e.g., a no-treatment control group) of interest to the researcher, we discuss the uses and limitations of an effect size derived from the SMD between pre-treatment and post-treatment

Correspondence concerning this article should be addressed to William T. Hoyt, University of Wisconsin-Madison, 335 Education Building, 1000 Bascom Mall, Madison, WI 53706-1398, USA. Email: wthoyt@wisc.edu; wthoyt@education.wisc.edu

scores for the treatment group (i.e., use of a within-group SMD when data are not available to compute a between-group SMD). We also discuss pros and cons of effect sizes which take baseline status (i.e., pre-treatment scores on the outcome variable) into account. For studies involving dichotomous outcomes (e.g., alive versus dead; employed versus not; clinically significant improvement versus not), readers are advised to compute an odds ratio to quantify differential rates of success for the two treatment conditions. Meta-analysis of odds ratios is beyond the scope of the present article. When all studies report on dichotomous outcomes, we refer readers to other sources (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Fleiss & Berlin, 2009) for assistance with effect size computation and analyses. When most studies report on mean differences for continuous outcomes, but a small subset of studies reports on conceptually parallel dichotomous outcome variables, procedures for computing odds ratios and converting to SMDs are available (Borenstein et al., 2009, Ch. 7).

Computing SMDs

The basic formula for Cohen's *d*, which estimates the SMD between two groups, requires data on the mean and standard deviation of the outcome scores for each group:

$$d = \frac{M_T - M_C}{S}, \tag{1}$$

where M_T and M_C are the means for treatment and comparison groups, respectively, and S is an estimate of the population standard deviation (σ) on the outcome variable. Below we consider several options for estimating σ. In meta-analysis, effect sizes are weighted by the (inverse of the) *sampling variance* (V), which reflects how precisely d estimates the population SMD. The variance of d is computed as

$$V_d = \frac{1}{n_T} + \frac{1}{n_C} + \frac{0.5d^2}{n_T + n_C}, \tag{2}$$

where n_T and n_C are the sample sizes for treatment and comparison groups, respectively, and d is computed by Equation (1).

Small-sample Bias in *d*

Because d is a slightly biased estimator of the population SMD when sample sizes are small, Hedges (1981; Hedges & Olkin, 1985) recommended computing a bias-corrected SMD estimator that has come to be called Hedges' g. The conversion from d to g is accomplished using a bias correction factor:

$$\mathcal{J} = 1 - \frac{3}{4 \cdot df - 1}, \tag{3}$$

which is then used to derive unbiased estimates of the SMD and its sampling variance[1]:

$$g = \mathcal{J} \cdot d, \% \tag{4a}$$

$$V_g = \mathcal{J}^2 \cdot V_d.\% \tag{4b}$$

What If the Primary Study Does Not Report *M*s and SDs?

A more challenging situation arises when means and standard deviations are not included in the primary research report. Often, though, Cohen's d (and hence, Hedges' g) can be computed from other information that is reported, such as a t statistic, an F statistic (1 degree of freedom in the numerator), or the exact p value from such a t- or F-test. Formulas to compute d given this information are well known, and are summarized in Borenstein (2009; Table 12.1).

When the research report does not provide specific enough information to compute a sample d for the study, the meta-analyst should contact the study's first author in an effort to obtain the needed data. Failing this, the meta-analyst must use judgment to determine how to proceed. An example of such a dilemma is a study that finds no significant difference between group means, but fails to report information (such as a t statistic or exact p value) needed to estimate what the SMD was in the sample. Omitting such a study from the meta-analytic dataset is not a good option, as this would systematically exclude studies that obtained smaller effect sizes, leading to a biased meta-analytic "sample." Instead, the preferred approach is to include this study in the dataset, and make the conservative assumption that $d = 0$. (Because this practice produces a *negative* bias in the synthesized effect estimates, it is a good idea to analyze data both with and without these studies, and discuss in a footnote when the two sets of results differ appreciably.)

Similarly, when group means (M, SD) are omitted from the research report, test statistics are not reported, and the exact p value is not reported, authors sometimes publish a significance level (e.g., "$p < .05$") on the basis of which they concluded that groups differed significantly. When this is the only information available, computing d under the

assumption that the observed p value was equal to the significance level (e.g., $p = .05$) produces a conservative estimate of the obtained effect size. As a final example, although APA requires that when means are reported authors should also report SDs, it is sometimes the case in non-APA journals (or in dissertations or other non-journal publications), the only effect size-related information provided in a research report are the group means. This allows for the computation of the raw mean difference, but provides no way to compute S (the estimate of the population standard deviation on the outcome variable) in Equation (1). In such a circumstance, the best option may be to look for other published data on the measure in question to obtain a reasonable value for S, with a preference for published studies with a sample as similar as possible to that in the incomplete research report.

In summary, in a meta-analysis of group differences, it is likely that the vast majority of research reports will provide sufficient information (group Ms and SDs; test statistics; exact p values) to compute an estimate of the population SMD from sample data. For the remaining studies, the meta-analyst must make judicious use of the data that are provided. When the choice is between excluding a study for incomplete data and making a conservative assumption to allow computation of an effect size based on the data provided, the latter option is preferred, to avoid positively biasing meta-analytic findings.

How Should I Estimate σ?

In Equation (1), S is an estimate of the population standard deviation (σ) on the outcome variable, generally computed from sample data. Many studies will report group Ms and SDs, and for these studies Hedges and Olkin (1985) considered two possibilities for computing S. Option 1 is to "pool" data from the two group SDs. The pooled SD is the square root of the weighted average of the group variances (i.e., SD_T^2 and SD_C^2), where each group is weighted by its degrees of freedom (i.e., group $n - 1$). The formula for the pooled SD is readily available (see, e.g., Borenstein et al., 2009, Equation (4.19)). Option 2 is to use the SD for the control group, on the grounds that treatments may affect the SD as well as the mean on the outcome variable, in which case the treatment group SD would be a biased estimate of the population SD. On statistical grounds, Hedges and Olkin recommended Option 1, as this yields a more precise estimator (S), which in turn leads to a more robust SMD estimate (d, g).

Alternative SMD Estimators: Variations in Experimental Design Among Studies

Carlson and Schmidt (1999) considered variants in the procedure used to estimate the SMD based on the design of the primary study. Three main experimental designs are used in program evaluation research. Equation (1) presumes a basic experimental design in which a control (or comparison) group is included, and outcomes are assessed following the intervention. Carlson and Schmidt refer to this design as *post-test only with control* (POWC) and to the estimator shown in Equation (1) as d_{POWC}. An alternative design, very common in psychotherapy outcome research, includes an assessment of the main outcome(s) at pre-treatment as well as post-treatment: *pre–post with control* (PPWC). This design allows for an alternative approach to estimate the SMD (d_{PPWC}), which corrects the post-treatment mean difference for any differences between groups that were present at baseline:

$$d_{PPWC} = \frac{(T_{post} - T_{pre}) - (C_{post} - C_{pre})}{S_{pre}}, \qquad (5)$$

where S_{pre} is the pooled SD for the two groups at baseline. Carlson and Schmidt (1999) considered d_{PPWC} to be the gold standard, as it corrects for pre-existing group differences, which are present even when participants are randomly assigned to groups, especially when sample sizes are not large (Hsu, 1989).

A third design is the less desirable *single group pre–post* (SGPP) design, which lacks a control group. When a study lacks a control group, the SMD can only be estimated as an index of pre–post change in the treatment group:

$$d_{SGPP} = \frac{T_{post} - T_{pre}}{SD_{pre}}. \qquad (6)$$

SGPP is deprecated in the experimental design literature, because the lack of a control group creates inferential problems: The change in the outcome variable over the course of treatment may be attributable to causes (such as maturation or history) other than the intervention (Shadish, Cook, & Campbell, 2002), and there is no control for these potential confounds. However, as noted by Carlson and Schmidt (1999), the SGPP design may be the only practical approach to program evaluation in some treatment contexts, and it is important to understand the limitations of d_{SGPP} as an SMD estimator, so that meta-analysts can make an informed decision about whether such studies can provide an unbiased estimate of treatment effects.[2] Using d_{SGPP} can also be

the only option for meta-analyses of psychotherapy outcomes when studies include a comparison group other than the control group of interest. For example, in a meta-analysis of treatments compared with no-treatment controls, the meta-analysis must determine whether and how to include a well-conducted randomized controlled trial that compares the treatment of interest to a placebo control group, or to an alternative treatment condition, but does not include a wait list control group that would permit computation of d_{POWC} or d_{PPWC}.

In this section, we offer considerations for combining effect sizes (d_{SGPP}, d_{POWC}, d_{PPWC}) based on different research designs, focusing first on the question of what can be done about studies that lack an appropriate control group, and next on whether and how baseline data should be included in estimating the SMD.

Is It Appropriate to Make Use of Data From Studies That Lack a Control Group?

There are two reasons to consider excluding data from studies lacking a control group. First, it can be argued that these studies are of poor methodological quality and therefore unlikely to provide valid data on the research question of interest. Second (and related), it may be that effect sizes computed from these studies (i.e., d_{SGPP}) are systematically biased. For example, in a meta-analysis of bereavement interventions (Hoyt, Del Re, & Larson, 2011), it is expected that symptoms will ameliorate over time to some extent in the absence of treatment. Studies that compare outcomes for a treatment group to those for a no-treatment group control for this natural improvement, and d_{POWC} (or d_{PPWC}) estimates improvement for the treatment group beyond the expected improvement in the absence of treatment. For studies lacking a no-treatment control group, d_{SGPP} will be a positively biased estimate of this treatment effect, because it includes gains due to treatment and also those that are expected in the absence of treatment.

The first objection to inclusion of no-control studies in a meta-analysis (namely, that these studies are of poor quality) can be challenged. Studies that lack a control group may provide data on innovative interventions or rare populations for which controlled trials are not yet available, but which would enhance generalizability of meta-analytic findings. Also, some psychotherapy trials are of high methodological quality but exclude a no-treatment control condition, either for ethical reasons or on the grounds that the efficacy of psychotherapeutic interventions relative to no-treatment controls is well established, and clinical trial resources are better devoted to other comparison conditions (e.g., treatment-as-usual; placebo controls; alternative treatments). So it may be problematic to equate lack of a control group to poor methodological quality, and there may be reasons for wishing to include studies of a range of methodological quality in a meta-analytic review.

The second objection (bias in estimating the SMD) is more compelling. Lipsey and Wilson (1993) published a meta-analysis of meta-analytic findings in psychological and educational research. They noted that, for those meta-analyses that directly compared effect sizes derived from pre–post only designs (d_{SGPP}) with those from studies that included a control group (usually d_{POWC}), the average difference in effect size was $d_{diff} = .29$, favoring the pre–post only studies. This suggests that d_{SGPP} may be a positively biased estimator of the population SMD, and that meta-analysts should be cautious about including data from single group designs unless this potential bias is examined directly and adjusted for. In the next section, we describe a method proposed by Becker (1988) that may allow investigators to include data from studies for which only d_{SGPP} can be computed, without biasing the findings of the meta-analysis.

Is It Important to Use pre-treatment Data When Estimating the SMD?

Despite the fact that most clinical trials report group means both pre- and post-treatment, the vast majority of meta-analyses of psychotherapy outcomes use Equation (1) to estimate the SMD for each study. This approach ignores pre-treatment data, computing d_{POWC} (in Carlson & Schmidt's, 1999 notation) even though the experimental design of the study is PPWC. Following Carlson and Schmidt, we might prefer computing d_{PPWC} (Equation (6)) for two reasons. First, compared with d_{POWC}, d_{PPWC} corrects for failures of randomization (i.e., baseline differences between groups on the outcome variable). Because these baseline differences are random (assuming random assignment was used), they will tend to cancel one another out across studies, and should not bias the omnibus effect size estimate derived from averaging d_{POWC} estimates rather than d_{PPWC} estimates.

Carlson and Schmidt (1999) considered another advantage of d_{PPWC} relative to d_{POWC}: namely, d_{PPWC} estimates S (in the denominator of the formula for all SMD estimators) using the pooled

pre-treatment standard deviation (notated as S_{pre} in Equation (5)) whereas d_{POWC} necessarily relies on post-treatment SDs. Glass, McGaw, & Smith (1981) recommended the use of pre-treatment data, when available, for estimating the population σ, because it is reasonable to expect that treatment interventions have an effect on the variability as well as the mean of the outcome scores (e.g., because of interactions between baseline status and treatments). In their dataset, derived from $k = 248$ studies of training interventions using the PPWC design, Carlson and Schmidt examined the hypothesis that post-treatment *SD*s are larger than pre-treatment *SD*s and found an average ratio for treated groups of $SD_{post}/SD_{pre} = 1.076$ (i.e., SD estimates were 7.6% larger post-treatment than at baseline), although the difference was not statistically significant. To demonstrate the effect on SMD estimates, they computed both d_{POWC} and d_{PPWC} for all studies, showing that d_{POWC} yielded an estimated SMD 4.3% smaller than d_{PPWC}. Thus, meta-analysts should prefer d_{PPWC} to d_{POWC} as an estimator of the population SMD when there is evidence in the meta-analytic dataset that post-treatment SDs are systematically different from pre-treatment SDs (suggesting treatment effects on variability as well as central tendency of the outcome variable).

Becker (1988) suggested an additional motivation for incorporating baseline data in effect size computation, in her development of a method to synthesize standardized mean-change scores. Becker sought to address the problem discussed in the previous section, that an appreciable proportion of studies in any meta-analysis use the SGPP design, and effect sizes from these studies (d_{SGPP}) may be biased because of failure to control for expected improvement in the absence of treatment. The norm in meta-analysis is to attempt to synthesize all available literature on the research question; so if a method is available that allows for computation of unbiased estimates of the SMD, even from these studies that lack a control group, then it is desirable to include them in the dataset.

Becker's (1988) recommended approach is to compute separate, within-group effect sizes (i.e., d_{SGPP}) for both treatment and (as available) control groups in each study. These standardized mean-change scores (assessing pre–post change in SD units for each group) are then corrected for small-sample bias (similar to Equation (4)) to produce standardized estimates of pre–post change in each group (we refer to these as g_C and g_T) and their variances (V_{g_C} and V_{g_T}).[3] Studies using the PPWC design produce estimates for both groups, and an effect size comparing change in the treatment group to change in the control group (Becker's Δ) is computed as the difference between the two within-group change scores:

$$\Delta = g_T - g_C, \tag{7}$$

and

$$V_\Delta = V_{g_T} + V_{g_C}. \tag{8}$$

Becker's Δ is conceptually and computationally similar to Carlson and Schmidt's d_{PPWC}. Both indices quantify change in the treatment group, adjusted for change in the control group, for each PPWC study in the dataset.

Becker's (1988) method of computing separate within-group effect sizes also allows for separate meta-analysis of change in treatment and control groups. This analysis is informative in its own right, and Becker suggested that the meta-analysis of control group change has an additional advantage, in that it allows for the imputation of control group effect sizes to (SGPP) studies that lack a control group. Although it was developed 30 years ago, Becker's method has been slow to gain popularity in published meta-analyses. Because of the prevalence of the PPWC experimental design in the psychotherapy outcome literature, this is an area where Becker's Δ may be especially attractive, with the benefit of allowing for inclusion in the dataset of both studies conducted using the SGPP design and those that used experimental designs but did not include the comparison condition of interest to the meta-analyst (see, e.g., Wade, Hoyt, Kidwell, & Worthington, 2014).

What If I Need to Use Different Estimators for Different Studies?

The procedure used to estimate the SMD for each study will necessarily be based on the experimental design of the study and the data provided in the research report. So it is usual that some studies (often the majority) provide data to compute d_{PPWC} (or alternatively Becker's Δ), whereas it will be natural to compute d_{POWC} or even d_{SGPP} for other studies in the meta-analysis. If the assumption of homogeneity of variance (pre- and post-treatment) holds, then d_{PPWC} and d_{POWC} are estimates of the same population SMD, the difference in outcomes between those in the treatment group and those in the comparison group. So there is no objection to combining these estimates across studies, even though the computational procedures differ. But d_{SGPP} estimates a different population SMD: the standardized mean change for the treatment group.

And there is good reason to believe, both theoretically (because symptoms may remit somewhat in the absence of treatment) and empirically (e.g., Lipsey & Wilson, 1993) that the SMD for absolute change in the treatment group differs from that for relative change in the treatment group (compared with controls). Thus, in the absence of some statistical adjustment (such as the procedures recommended by Becker, 1988, for imputation of control group change effect sizes), it is problematic to combine d_{SGPP} effect sizes from some studies with d_{PPWC} or d_{POWC} effect sizes from other studies.

Computing Multiple Effect Sizes from a Single Study

When research reports allow for computation of multiple effect sizes related to the research question, as is common, meta-analysts need to be aware of the issue of statistical dependency of effect sizes. We discuss this issue in more detail in the next section. Here we introduce three common features of studies that allow for computation of multiple effect sizes: multiple treatment or comparison groups, multiple endpoints, and multiple outcome measures.

Multiple Groups

When an experimental design involves more than two groups, this allows for multiple pairwise comparisons among those groups. Each of these pairwise comparisons can be quantified as an effect size (d_{PPWC} or d_{POWC}). When a study includes one treatment condition and multiple comparison conditions (e.g., wait list control; placebo control), it is often the case that only one of the possible effect sizes belongs in the meta-analysis, in which case this design does not pose difficulties with dependent effect sizes. Comparison to a wait list control and comparison to a placebo control address different research questions, so it is likely that the two effect sizes addressing these questions would be included in different meta-analyses.

When a study includes two or more treatment conditions and a single comparison condition, the meta-analyst must first determine whether both treatments meet the inclusion criteria for the meta-analysis. If so, the simplest method to combine the two effect sizes is to combine the data from the two treatment conditions, using the weighted mean (weighted by group size) and the pooled standard deviation as *M* and SD for the combined treatment condition. (If the two treatments have different values on some moderator variable, one can of course disaggregate the effect sizes for the moderator analysis; Cooper, 1998.)

Multiple Endpoints

In addition to assessing outcomes immediately following the end of treatment, it is common for clinical trials to include one or more follow-up assessments, to examine maintenance of gains for the principal outcomes. Meta-analysts can and should compute an effect size for each of the endpoints. Generally, the post-treatment effect sizes are examined in a separate analysis than the follow-up effect sizes, which avoids much of the potential for dependency arising from this design feature. For a meta-analysis of follow-up outcomes, when one study includes multiple follow-ups, the issue of dependency can arise, and investigators often resolve this issue by including only the effect size for the longest follow-up.

Multiple Outcomes

It is also common for clinical trials to include data for several outcomes. Primary researchers may include multiple measures of the focal outcome variable (e.g., several measures of depression symptoms, one client-rated and one rated by an expert interviewer). Or they may assess the focal outcome and one or more peripheral outcomes that are expected to improve as a result of symptom amelioration (e.g., quality of social relationships; work functioning; global quality of life). If the relevant data are provided in the research report, the meta-analyst can compute an effect size for each outcome assessed, and then must determine the appropriate approach to inclusion of these data in the analysis.

One option is to conduct separate meta-analyses for several theoretically important outcome variables. This addresses questions about treatment effects on specific outcomes, and allows for examination of relative efficacy for reduction of particular symptoms. (However, because not all studies contribute effect sizes for every outcome, it is advisable to conduct statistical tests of relative efficacy using a within-studies approach, to avoid confounding with between-study differences in overall efficacy.)[4] This targeted approach likely also means ignoring data from some outcomes for many of the studies, to focus on just a subset of the outcomes assessed.

Smith and Glass (1977) argued for an alternative approach – an omnibus analysis incorporating data from all constructs assessed in each study, followed (if desired) by targeted follow-up analyses for specific outcomes. Their rationale for the omnibus analysis is

that psychological outcomes are generally correlated, and we should trust that primary investigators are judicious in their selection of outcome variables that are (a) relevant to the presenting issue and (b) likely to improve as a result of treatment. Smith and Glass noted that, from a policy perspective, one can imagine posing the question "What kind of effect does therapy produce – on anything" (p. 753). The omnibus analysis, including all constructs measured in each study, arguably best addresses this global question about effectiveness. A counter-argument to this approach is that effect sizes in the omnibus analysis are confounded by differences in study measurement procedures. Studies that measure only targeted outcomes will likely have larger effect sizes, and studies that measure many peripheral outcomes will have smaller aggregate effect sizes, when the effects are averaged across all outcomes for the omnibus analysis. So the effects in the omnibus analysis may reflect researcher choices about what constructs to assess, in addition to actual treatment effects.

Whichever approach is adopted, it is likely that at least some studies will provide multiple estimates of the treatment effect, and these multiple effect sizes drawn from the same sample are not statistically independent. Meta-analysts will need to consider how to address this problem of statistical dependency, which is the topic of the next section.

Statistical Dependence Among Effect Sizes

When a single study provides two or more effect estimates relevant to the research question of interest, these effect sizes will be statistically dependent. If they are included separately in the dataset to be analyzed, they will be treated in the meta-analysis as statistically independent observations of the treatment effect, resulting in improper effect size weighting and biased significance tests and confidence intervals (Gleser & Olkin, 2009). To avoid this undesirable result, meta-analysts have long been advised to aggregate dependent effect sizes prior to analysis (e.g., Cooper, 1998; Lipsey & Wilson, 2001), so that each study contributes only a single effect size to each combined effect estimate within the meta-analysis. The methods we describe here are for aggregation of Cohen's *d*. Users should compute *d* for each comparison available in each study, aggregate these to create a composite effect size, then use the procedures in Equations (3) and (4) to convert the composite *d* (and V_d) for each study to the unbiased estimator *g* (and V_g).

Some care must be taken in using aggregation procedures that will provide a proper estimate of sampling variance for the resulting composite effect size. Two statistically rigorous methods have been proposed, both requiring knowledge of the covariances (correlations) among dependent effect sizes. Hedges and Olkin (1985) proposed that the composite effect size should be a weighted mean of the dependent effect sizes, taking intercorrelations among the effect sizes into account, and showed how to compute the sampling variance for this estimator (see also Gleser & Olkin, 1994). More recently, Borenstein et al. (2009, Ch. 24) proposed that the composite effect can be computed straightforwardly as the unweighted mean of the dependent effect sizes (Borenstein et al., 2009, Equation (24.4)), and showed how to compute the sampling variance for this estimator (Equation (24.5)):

$$\bar{d} = \frac{1}{m}\left(\sum_{j=1}^{m} d_j\right), \tag{9}$$

and

$$V_{\bar{d}} = \left(\frac{1}{m}\right)^2 \left(\sum_{j=1}^{m} V_j + \sum_{j \neq k} r_{jk}\sqrt{V_j}\sqrt{V_k}\right), \tag{10}$$

where $\bar{d}$ is the composite effect size, $V_{\bar{d}}$ is its variance, and *m* is the number of effect sizes to be aggregated for the study. In a simulation study (Hoyt & Del Re, n.d.), our statistical results slightly favored the Borenstein et al. approach, which is also easier to use; so this is the approach we recommend. This aggregation approach has been implemented in the Comprehensive Meta-Analysis software package (Borenstein, Hedges, Higgins, & Rothstein, 2013), and also in the MAd package in R (Del Re & Hoyt, 2014).

We note also that some meta-analysts (e.g., Scammacca, Roberts, & Stuebing, 2014) have used a variant of the Borenstein et al. method: computing the composite effect size as the unweighted mean of the dependent effect sizes, but estimating the sampling variance as the unweighted mean of their variances. This method, which Hoyt and Del Re (n.d.) called the *naïve* approach, yields an inflated estimate of the sampling variance for the composite effect size. This results in penalizing studies that assess multiple outcomes by underweighting their information when synthesizing the effect sizes across studies (Borenstein et al., 2009, pp. 237–238; Hoyt & Del Re, n.d.). Thus, estimating the variance of the composite effect size as the unweighted mean of the variances of the dependent effect sizes is not recommended.[5]

What If the Primary Study Does Not Report the Correlations Among Outcome Measures?

Both of the statistically rigorous methods for aggregating dependent effect sizes described in the previous section require estimates of the population correlations (ρ) between pairs of outcome variables. The obvious place to seek this information is the research report itself, but it is relatively unusual for authors to provide this information. Thus, other sources may need to be consulted for an estimate of the population correlation. If a different study can be located which included the same two measures (ideally in a sample from a similar population, and with a relatively large sample size), this can provide a robust estimate of the needed correlation coefficient. Failing this, one can use meta-analytic evidence about the relation between the two constructs (e.g., correlations between depression and anxiety) as a proxy estimate of ρ between specific measures of these constructs. (This was the basis for Wampold et al.'s (1997) decision to estimate the correlation between pairs of dependent outcomes as $\rho = .50$.) In fact, because sample sizes may be relatively small in clinical trials, relying on large sample or meta-analytic estimates is a sensible strategy, as these yield more precise estimates of the desired population correlation.

Researchers and reviewers are sometimes uneasy about the aggregation methods described because they worry that the estimated correlation may not be accurate, leading to a biased estimate of sampling variance for the composite effect size. It is true that misestimation of ρ will introduce a source of error in the meta-analytic effect estimates. But this will be small compared to the error introduced by use of the naïve method (which in effect imputes a correlation of $\rho = 0$ between outcome measures), or by including multiple dependent effect sizes in the meta-analytic dataset.

When there is uncertainty about the population correlation between outcome measures, Borenstein et al. (2009) recommend conducting a sensitivity analysis. For example, if the outcome correlation is estimated as $\rho = .50$, the meta-analyst could conduct separate aggregations and analyses at deviant ρ values (e.g., $\rho = .40$; $\rho = .60$) to determine whether these alternative estimates lead to a substantive difference in results or conclusions.

Other Approaches to Addressing Effect Size Dependency

Gleser and Olkin (2009) recommended an alternative approach to addressing statistical dependencies in meta-analysis via multilevel modeling. In this approach, all effect sizes are included in the dataset, and nesting of effect sizes within studies is specified at Level 2, with between-studies differences analyzed at Level 3. (Level 1 is the participant level, and, as participant data are lacking, the variance component for participants within effect sizes is specified for each study as the sampling variance for that effect size.) Cheung (2013) created an R package for conducting three-level meta-analysis in SEM, to assist with implementing this approach.

Hedges, Tipton, and Johnson (2010) developed a procedure for robust variance estimation (RVE) in the presence of dependent effect sizes that avoids the problem of having to estimate population correlations between pairs of outcome variables. (Population correlations are estimated in these models, but make little difference to the results.) Meta-analysts can include all effect sizes in the dataset and implement the RVE procedure to obtain an unbiased summary effect size and accurate standard error. Tanner-Smith and Tipton (2014) provided a tutorial for using the RVE method through use of either Stata or SPSS macros, and conducted simulation studies to identify boundary conditions on the use of this method. They note that the method performs well for estimating an average effect size in even relatively small meta-analyses (e.g., $k = 10$), but yields negatively biased estimates of the standard errors for meta-regression coefficients (i.e., slope coefficients in moderator analyses) when the number of studies is smaller than $k = 40$.

Summary

It is likely that meta-analysts studying psychotherapy outcomes will find that at least some (and usually many) of the studies provide data for computing multiple effect sizes related to the research question. In such cases, one must take care that analyses do not violate assumptions of independence of observations, using one of the approaches described here. The first approach (aggregation of effect sizes within studies so that each study contributes only a single independent effect size) is an attractive option because it yields unbiased effect sizes and standard errors, and uses basic statistical procedures. An additional advantage of computing an aggregate effect size for each study is that it makes possible tabular and graphical representations (e.g., forest plots; funnel plots) that provide basic descriptive information and address questions of publication bias in the meta-analytic dataset. We also described two alternative methods to address

dependency by specifying the nesting of effect sizes within studies.

Conclusions

Although the basic computation of effect sizes (SMDs) is relatively straightforward, it is common for researchers meta-analyzing data on intervention studies to encounter a variety of complications, including incomplete reporting of quantitative data, options for combining effect sizes from studies using different research designs (e.g., different comparison groups or no comparison group), and studies that yield multiple estimates of the effect of interest. In addition, meta-analysts have options about how to estimate σ (the population SD on which the mean difference is standardized) and whether to make use of baseline data in effect size computation (i.e., computing the difference in mean-change scores rather than post-treatment means). In this paper, we discussed guidelines for the many decisions in the process of effect size computation, with citations to resources that provide additional information. Meta-analysts should describe their procedures for computing and aggregating effect sizes clearly in the published report, so that readers have full information about the methods used.

Notes

[1] There has been a shift in meta-analytic notational conventions over time that can be confusing for researchers new to the literature on this method. Hedges and Olkin (1985) denoted the biased estimator (Equation (1)) as *g*, and the corrected estimator (Equation (4a)) as *d*. Much early meta-analytic work followed this notational convention. Starting with the second edition of the *Handbook of Research Synthesis* (Cooper, Hedges, & Valentine, 2009), the opposite notational convention has become standard: *d* (i.e., Cohen's *d*) denotes the basic formula (Equation (1)), which represents a slightly biased estimate for small-*N* studies; the bias-corrected estimator (Equation (4a)) is notated as *g*. We follow this modern notational convention throughout this article, and note where this deviates from the notation used in some of the earlier papers we cite.

[2] A slight complication in the use of d_{SGPP} involves the computation of the variance (V_d), which requires adjustment for the correlation between pre and post scores. Borenstein (2009) provides the formula for this variance (Equation (12.21)):

$$V_d = \left(\frac{1}{n} + \frac{d^2}{2n}\right) \cdot 2(1 - r).$$

Because the pre–post correlation (*r*, above) is often not reported, it is common to impute it. This is a type of test–retest correlation, but is likely to be attenuated because of differential response to intervention, so $r = .5$ or .6 may be a reasonable estimate.

[3] Becker (1988) followed the original Hedges and Olkin's (1985) notational convention (i.e., *g* was the biased estimator and *d* was the bias-corrected estimator of standardized mean-change). We have updated the notation (see Endnote 1); so our formulas do not match those in the original article (i.e., *g* and *d* are exchanged in our version of Becker's formulas).

[4] A within-studies analysis focuses on just those studies that provide an effect size for both constructs. For example, to compare the effects of cognitive-behavioral treatments on depression symptoms and anxiety symptoms, one would select the subset of studies that measure both depression and anxiety, compute *g* and V_g for each outcome, and then compute the difference between these two (i.e., $g_{diff} = g_{depr} - g_{anx}$) and its variance ($V_{g_{diff}} = V_{g_{depr}} + V_{g_{anx}} - 2r\sqrt{V_{g_{depr}} V_{g_{anx}}}$), where *r* is the assumed correlation between depression scores and anxiety scores). One then meta-analyzes g_{diff}, and if the 95% CI for this difference score excludes zero, this supports the hypothesis that the treatment has different effects on depression and anxiety (cf. Borenstein et al., 2009, pp. 233–235).

[5] Use of improper aggregation methods in meta-analysis results in error rather than bias in meta-analytic results. That is, the improper aggregation may result in an effect size that is either higher or lower than that based on optimal aggregation procedures, depending on the size of the effect estimates in studies that include multiple outcome measures, relative to those that assess only a single outcome. Improper aggregation procedures also result in inaccurate results for moderator tests, again due to incorrect weighting of studies in the analysis.

References

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257–278.

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (2nd ed., pp. 221–235). New York, NY: Russell Sage.

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. (2013). *Comprenensive meta-analysis* (computer software), *Version 3*. Englewood, NJ: Biostat.

Carlson, K. D., & Schmidt, F. L. (1999). Impact of experimental design on effect size: Findings from the research literature on training. *Journal of Applied Psychology*, *84*(6), 851–862. doi.org/10.1037//0021-9010.84.6.851

Cheung, M. W.-L. (2013). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*. doi.org/10.1037/a0032968

Cooper, H. (1998). *Synthesizing research: A guide for literature reviews* (3rd ed.). Thousand Oaks, CA: Sage.

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 3–16). New York, NY: Russell Sage.

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

Del Re, A. C., & Hoyt, W. T. (2014). *MAd: Meta-analysis with mean differences*. R-package version 0.8-2 (computer software). Retrieved from http://cran.r-project.org/web/packages/MAd

Fleiss, J. L, & Berlin, J. A. (2009). Effect sizes for dichotomous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.),

The handbook of research synthesis and meta-analysis (2nd ed, pp. 237–253). New York, NY: Russell Sage.

Glass, G. V, McGaw, B, & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (1st ed., pp. 339–355). New York, NY: Russell Sage.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *Handbook of research synthesis and meta-analysis* (2nd ed, pp. 357–376). New York, NY: Russell Sage.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis* (Vol. 20). New York, NY: Academic Press.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. doi.org/10.1002/jrsm.5

Hoyt, W. T., & Del Re, A. C. (n.d.). Comparison of methods for aggregating dependent effect sizes in meta-analysis.

Hoyt, W. T., Del Re, A. C., & Larson, D. G. (2011). *A new look at the evidence: Grief counseling is effective*. 9th International Conference on Grief and Bereavement in Contemporary Society, Miami, FL.

Hsu, L. M. (1989). Random sampling, randomization, and equivalence of contrasted groups in psychotherapy outcome research. *Journal of Consulting and Clinical Psychology*, *57*(1), 131–137. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2647799

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. Confirmation from meta-analysis. *The American Psychologist*, *48*(12), 1181–1209. doi.org/10.1037//0003-066X.48.12.1181

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research*, *84*, 328–364.

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

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. doi.org/10.1037/0003-066x.32.9.752

Tanner-Smith, E, & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*, 13–30.

Wade, N. G., Hoyt, W. T., Kidwell, J. E. M., & Worthington, E. L. (2014). Efficacy of psychotherapeutic interventions to promote forgiveness: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *82*(1), 154–170. doi.org/10.1037/a0035268

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "All must have prizes". *Psychological Bulletin*, *122*(3), 203–215.