Meta-Psychology, 2020, vol 4, MP.2018.884,
https://doi.org/10.15626/MP.2018.884
Article type: Original Article

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Daniël Lakens
Reviewed by: K. Coburn, F. D. Schönbrodt
Analysis reproduced by: Erin M. Buchanan
All supplementary files can be accessed at the OSF project page: https://doi.org/10.17605/OSF.IO/MY8S6

Publication Bias in Meta-Analyses of Posttraumatic Stress Disorder Interventions

Helen Niemeyer*
Freie Universität Berlin, Department of Clinical Psychological Intervention, Germany

Robbie C.M. van Aert*
Tilburg University, Department of Methodology and Statistics, the Netherlands

Sebastian Schmid
Department of Psychiatry and Psychotherapy, Charité University Medicine Berlin, Campus Mitte, Germany

Dominik Uelsmann
Department of Psychiatry and Psychotherapy, Charité University Medicine Berlin, Campus Mitte, Germany

Christine Knaevelsrud
Freie Universität Berlin, Department of Clinical Psychological Intervention, Germany

Olaf Schulte-Herbrueggen
Department of Psychiatry and Psychotherapy, Charité University Medicine Berlin, Campus Mitte, Germany

* HN and RvA have equally contributed and split first authorship.

Meta-analyses are susceptible to publication bias, the selective publication of studies with statistically significant results. If publication bias is present in psychotherapy research, the efficacy of interventions will likely be overestimated. This study has two aims: (1) investigate whether the application of publication bias methods is warranted in psychotherapy research on posttraumatic stress disorder (PTSD) and (2) investigate the degree and impact of publication bias in meta-analyses of the efficacy of psychotherapeutic treatment for PTSD. A comprehensive literature search was conducted and 26 meta-analyses were eligible for bias assessment. A Monte-Carlo simulation study closely resembling characteristics of the included meta-analyses revealed that statistical power of publication bias tests was generally low. Our results showed that publication bias tests had low statistical power and yielded imprecise estimates corrected for publication bias due to characteristics of the data. We recommend to assess publication bias using multiple publication bias methods, but only include methods that show acceptable performance in a method performance check that researchers first have to conduct themselves.

Keywords: Publication bias, meta-meta-analysis, meta-analysis, posttraumatic stress disorder, psychotherapy

Meta-Psychology, 2020, vol 4, MP.2018.884,
https://doi.org/10.15626/MP.2018.884
Article type: Original Article

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Daniël Lakens
Reviewed by: K. Coburn, F. D. Schönbrodt
Analysis reproduced by: Erin M. Buchanan
All supplementary files can be accessed at the OSF project page: https://doi.org/10.17605/OSF.IO/MY8S6

Posttraumatic stress disorder (PTSD) following potentially traumatic events is a highly distressing and common condition, with lifetime prevalence rates in the adult population of 11.7% for women and 4% for men in the United States of America (Kessler, Petukhova, Sampson, Zaslavsky, & Wittchen, 2012). PTSD is characterized by the re-experiencing of a traumatic event, avoidance of stimuli that could trigger traumatic memories, negative cognitions and mood, and alterations in arousal and reactivity (American Psychiatric Association, 2013). The DSM criteria have been updated recently, but most research is still based on the previous versions DSM-IV-TR (American Psychiatric Association, 2000), DSM-IV (American Psychiatric Association, 1994) or DSM-III-R (American Psychiatric Association, 1987).

Various forms of psychological interventions for treating PTSD have been investigated in a large number of studies. Cognitive behavioral therapies (CBT) and eye movement desensitization and reprocessing (EMDR) are the most frequently studied approaches (e.g., Bisson, Roberts, Andrew, Cooper, & Lewis, 2013). Trauma-focused cognitive behavioral therapies (TF-CBT) use exposure to trauma memory or reminders and the identification and modification of maladaptive cognitive distortions related to the trauma in their treatment protocols (e.g., Ehlers, Clark, Hackmann, McManus, & Fennell, 2005; Foa & Rothbaum, 1998; Resick & Schnicke, 1993). Non trauma-focused cognitive behavioral therapies (non TF-CBT) do not focus on trauma memory or meaning, but for example on stress management (Veronen & Kilpatrick, 1983). EMDR includes an imaginal confrontation of traumatic images, the use of eye movements and some core elements of TF-CBT (see Forbes et al., 2010). Although a range of other psychological treatments exists (e.g., psychodynamic therapies or hypnotherapy), fewer empirical studies of these approaches have been conducted (Bisson et al., 2013).

Meta-analysis methods are used to quantitatively synthesize the results of different studies on the same research question. Meta-analysis has become more popular according to the gradual increase of published papers that apply meta-analysis methods especially since the beginning of the 21st century (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2010), and it has been called the "gold standard" for synthesizing individual study results (Aguinis, Gottfredson, & Wright, 2011; Head, Holman, Lanfear, Kahn, & Jennions, 2015). Results of meta-analyses are often used for deciding which treatment should be applied in clinical practice, and international evidence-based guidelines recommend TF-CBT and EMDR for the treatment of PTSD (ACPMH; Forbes et al., 2007; NICE; National Collaborating Centre for Mental Health, 2005).

Publication Bias in Psychotherapy Research

The validity of meta-analyses is highly dependent on the quality of the included data from primary studies (Valentine, 2009). One of the most severe threats to the validity of a meta-analysis is publication bias, which is the selective reporting of statistically significant results (Rothstein, Sutton, & Borenstein, 2005). Approximately 90% of the main hypotheses of published studies within psychology are statistically significant (Fanelli, 2012; Sterling, Rosenbaum, & Weinkam, 1995) and this is not in line with the on average low statistical power of studies (Bakker, van Dijk, & Wicherts, 2012; Ellis, 2010). If only published studies are included in a meta-analysis, the efficacy of interventions may be overestimated (Hopewell, Clarke, & Mallett, 2005; Ioannidis, 2008; Lipsey & Wilson, 2001; Rothstein et al., 2005). About one out of four funded studies examining the efficacy of a psychological treatment for depression did not result in a publication, and adding the results of the retrieved unpublished studies lowered the mean effect estimate by 25% from a medium to a small effect size (Driessen, Hollon, Bockting, & Cuijpers, 2017).

The treatments in evidence-based psychotherapy are mainly selected based on published research (Gilbody & Song, 2000). The scientist-practitioner model (Shapiro & Forrest, 2001) calls for clinical psychologists to let empirical results guide their work, aiming to move away from opinion- and experience-driven therapeutic decision making toward the use of research results in clinical practice. If publication bias is present, guidelines may offer recommendations seemingly based on apparent empirical evidence that are only erroneously supported by the results of meta-analyses (Berlin & Ghersi, 2005).

Consequently, psychotherapists who follow the scientist-practitioner model would be prompted to apply interventions in routine care that may be less efficacious than assumed and may even have detrimental effects for patients.

A re-analysis of meta-analyses in psychotherapy research for schizophrenia and depression revealed that evidence for publication bias was found in about 15% of these meta-analyses (Niemeyer, Musch, & Pietrowsky, 2012, 2013). However, until now no further comprehensive assessment of publication bias in meta-analyses of the efficacy of psychotherapeutic treatments for other clinical disorders has been conducted. Hence, the presence and impact of publication bias in psychotherapy research also for PTSD remains largely unknown. Although trauma-focused interventions are claimed to be efficacious, their efficacy may be overestimated and might be lower if publication bias was taken into account. This in turn would result in suboptimal recommendations in the treatment guidelines and consequently also in unnecessarily high costs for the health care system (Jaycox & Foa, 1999; Maljanen et al., 2016; Margraf, 2009).

Due to publication bias being widespread and its detrimental impact on the results of meta-analyses (Dickersin, 2005; Fanelli, 2012; Rothstein & Hopewell, 2009), a statistical assessment of publication bias should be conducted in every meta-analysis investigating psychotherapeutic treatments. This is in line with recommendations in the Meta-Analysis Reporting Standards (MARS; American Psychological Association, 2010) and the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher, Liberati, Tetzlaff, & Altman, 2009). A considerable number of statistical methods to investigate the presence and impact of publication bias have been developed in recent years. These methods should also be applied to already published meta-analyses in order to examine whether publication bias distorts the results (Banks, Kepes, & Banks, 2012; van Assen, van Aert, & Wicherts, 2015).

The development of publication bias methods and recommendations to apply these methods will likely yield a more routinely assessment of publication bias in meta-analyses. However, research has shown that publication bias tests generally suffer from low statistical power and especially if there are only a small number of studies included in a meta-analysis and publication bias is not extreme (Begg & Mazumdar, 1994; Egger, Smith, Schneider, & Minder, 1997; Renkewitz & Keiner, 2019; Sterne, Gavaghan, & Egger, 2000; van Assen, van Aert, & Wicherts, 2015). This raises the question whether routinely applying publication bias tests without taking into account characteristics of the meta-analysis, such as the number of included studies, is a good practice.

Objectives

The first goal of this paper is to study whether applying publication bias tests is warranted under conditions that are representative for published meta-analyses on PTSD treatments. Applying publication bias tests may not always be appropriate if, for example, statistical power of these tests is low caused by a small number of studies included in the meta-analysis. Hence, we study the statistical properties of publication bias tests by conducting a Monte-Carlo simulation study that closely resembles the meta-analyses on PTSD treatments.

The second goal of our study is to assess the severity of publication bias in the meta-analyses published on PTSD treatments. We will not interpret the results of the publication bias tests if it turns out that these tests have low statistical power. Regardless of these results, we will apply multiple methods to correct effect size for publication bias to the meta-analyses on PTSD treatments. Effect size estimates of these methods become less precise (wider confidence intervals), but they still provide relevant insights into whether the effect size estimate becomes closer to zero if publication bias is taken into account.

Method

Data Sources

We conducted a literature search following the search strategies recommended by Lipsey and Wilson (2001) to identify all meta-analyses published on PTSD treatments. We screened the databases PsycINFO, Psyndex, PubMed, and the Cochrane Database of Systematic Reviews for all published and unpublished meta-analyses in English or German up to 5th September 2015. The search combined terms indicative of meta-analyses or reviews and terms indicative of PTSD. The exact search terms were [(“metaana*” OR “meta-ana*” OR “review” OR

"Übersichtsarbeit") AND ("stress disorders, post traumatic" (MeSH) OR "post-trauma*" OR "post-trauma*" OR "posttraumatic stress disorder" OR "trauma*" OR "PTSD" OR "PTBS")].

In addition, a snowball search system was used for the identification of further potentially relevant studies by screening the reference lists of included articles and of conference programs from the field of PTSD and trauma as well as psychotherapy research (see https://osf.io/9b4df/ for more information). Experts in the field were contacted, but no additional meta-analyses were obtained. Meta-analyses were retrieved for further assessment if the title or abstract suggested that these dealt with a meta-analysis of psychotherapy for PTSD. If an abstract provided insufficient information, the respective article was examined in order not to miss a relevant meta-analysis.

Study Selection and Data Extraction

Meta-analyses were required to meet the following inclusion criteria: 1) a psychotherapeutic intervention was evaluated. Psychotherapy was defined as "the informed and intentional application of clinical methods and interpersonal stances derived from established psychological principles for the purpose of assisting people to modify their behaviors, cognitions, emotions, and/or other personal characteristics in directions that the participants deem desirable" (Norcross, 1990, p. 219). 2) The intervention aimed at reducing subclinical or clinical PTSD, according to diagnostic criteria for PTSD (e.g., using one of the versions of the DSM) or according to PTSD symptomatology as measured by a validated self-report or clinician measure in an adult population (i.e., aged 18 years and older). And 3) a summary effect size was provided. Both uncontrolled designs investigating changes in one group (within-subjects design) and multiple group comparisons (between-subjects design) were suitable for inclusion. Exclusion criteria were: 1) pooling of studies with various disorders, so that samples composed of other disorders along with PTSD were included in a meta-analysis and the effect sizes were combined to an overall effect estimate not restricted to the treatment of PTSD; and 2) the meta-analysis examined the efficacy of pharmacological treatment. Three independent raters (DU, HN, SSch) decided on the inclusion or exclusion of each meta-analysis upon preliminary reading of the abstract and discussed in the case of dissent.[1] We included a meta-analysis if it did not explicitly target children and adolescents, but minor hints for the inclusion of such studies were present. However, this was only suitable if it concerned individual studies in a meta-analysis, and if we found such hints only when thoroughly checking the list of references.

For conciseness, we use the term meta-analysis to refer to the article that was published and use the term data set for the effect sizes included in a meta-analysis. A meta-analysis can comprise more than one data set if, for instance, treatment efficacy was investigated for different outcomes, such as PTSD symptoms and depressive symptoms, or when the efficacy of two treatments (e.g., TF-CBT and EMDR) was investigated separately in the same meta-analysis. The term primary study is used to refer to the original study that was included in the meta-analysis. When a meta-analysis consists of multiple data sets, we included all data sets for which primary studies' effect sizes and a measure of their precision were provided or could be computed.

We tried to extract effect sizes and their precision of the primary studies from the meta-analysis. If the required data were not reported, we contacted the corresponding authors and re-analyzed the primary studies in order to obtain the data. Data were extracted independently by one author (SSch), cross-checked by a second reviewer (HN), and in case of deviations during the statistical calculations checked by two researchers (RvA, HN). All data sets for which the data were available and we could reproduce the average effect size reported in the meta-analysis ourselves were included. An absolute difference in average effect size larger than 0.1 was set as criterion for reproducibility. We labeled a data set as not reproducible if we could not reproduce the results based on the available data and description of the analyses after contacting the authors of a meta-analysis. Moreover, there were no restrictions with respect to the dependent variable. That is, all primary and secondary outcomes of the meta-analyses were suitable for inclusion. Primary outcomes in meta-analyses on PTSD are usually PTSD symptom score or clinical status, whereas secondary outcomes often vary (e.g. anxiety, depression, dropout, or other; see also Bisson, Roberts, Andrew, Cooper, & Lewis, 2013).

The objectives of our paper were to study whether applying publication bias tests is warranted in meta-analyses on the efficacy of psychotherapeutic treatment for PTSD and to assess the severity of publication bias in these meta-analyses. The majority of statistical methods to detect the presence of publication bias does not perform well if the true effect sizes are heterogeneous (e.g., Stanley & Doucouliagos, 2014; van Aert et al., 2016; van Assen et al., 2015), some are even recommended not to be used in this situation (Ioannidis, 2005). Hence, it was necessary to only include data sets where the proportion of variance that is caused by heterogeneity in true effect size as quantified by the I^2-statistic was smaller than 50%.

We excluded all data sets of a meta-analysis that included less than six studies, because publication bias tests suffer from low statistical power in case of a small number of studies in a meta-analysis and if severe publication bias is absent (Begg & Mazumdar, 1994; Sterne et al., 2000). Others recommend a minimum of 10 studies (Sterne et al., 2011), but we adopted a less strict criterion for two reasons. First, we want to study whether applying publication bias tests is warranted for conditions that are representative for published meta-analyses. Meta-analyses often contain less than 10 studies. For example, the median number of studies in meta-analyses published in the Cochrane Database of Systematic Reviews is 3 (Rhodes, Turner, & Higgins, 2015; Turner, Jackson, Wei, Thompson, & Higgins, 2015). Also the number of studies in meta-analyses for psychotherapy research is usually small. Meta-analyses on the efficacy of psychotherapy for schizophrenia (Niemeyer, Musch, & Pietrowsky, 2012) as well as depression (Niemeyer, Musch, & Pietrowsky, 2013) also applied a minimum of 6 studies as lower limit for the application of publication bias tests.

Second, more recently developed methods to correct effect size for publication bias can be used to estimate the effect size even if the number of studies in a meta-analysis is small. For example, a method that was developed for combining an original study and replication has shown that two studies can already be sufficient for accurately evaluating effect size (van Aert & van Assen, 2018). However, a consequence of applying publication bias methods to meta-analyses based on a small number of studies is that effect size estimates become less precise and corresponding confidence intervals wider (Stanley et al., 2017; van Assen et al., 2015).

Statistical Methods

Publication bias test. We assessed for the following publication bias tests whether it was warranted to apply these methods to the data sets in PTSD psychotherapy research: Egger's regression test (Egger et al., 1997), rank-correlation test (Begg & Mazumdar, 1994), Test of Excess Significance (Ioannidis & Trikalinos, 2007b), and p-uniform's publication bias test (van Assen et al., 2015). These methods were included, because these are commonly applied in meta-analyses (Egger's regression test and rank-correlation test) or outperformed existing methods in some situations (TES and p-uniform's publication bias test; Renkewitz & Keiner, 2019). It is important to note that Egger's regression test and the rank-correlation test were developed to test for small-study effects. Small-study effects refer to the tendency of smaller studies to go along with larger effect sizes. One of the causes of small-study effects is publication bias, but another cause is, for instance, heterogeneity in true effect size (see Egger et al., 1997, for a list of causes of small-study effects). The TES was also not specifically developed to test for publication bias, but examines whether the observed and expected number of statistically significant effect sizes in a meta-analysis are in line with each other (see https://osf.io/b9t7v/ for an elaborate overview of existing publication bias tests).

In order to investigate whether the application of the publication bias tests to the included data sets was warranted, we conducted a Monte-Carlo simulation study to examine the statistical power of the publication bias tests for the data sets. Data were generated in a way to stay as close as possible to the characteristics of the data sets. That is, the same number of effect sizes as in the data set as well as the same effect size measure were used for generating the data. The data were simulated under the fixed-effect (a.k.a. equal-effects) model, so effect sizes for each data set were sampled from a normal distribution with mean μ and variance equal to the "observed" squared standard errors. Statistically significant effect sizes based on a one-tailed test with α=.025 (to reflect common practice of testing a two-tailed hypothesis and only reporting results in the predicted direction) were always "published"

and included in a simulated meta-analysis. Publication bias restricted the "publication" of statistically nonsignificant effect sizes in a way that these effect sizes had a probability of 1-pub to be included in a simulated meta-analysis. Effect sizes were simulated till the included number of simulated effect sizes equaled the number of effect sizes in a data set.

We examined the Type-I error rate and statistical power of Egger's regression test, rank-correlation test, TES, and p-uniform's publication bias test for each simulated meta-analysis using α=.05. Two-tailed hypothesis tests were conducted for Egger's regression test and the rank-correlation test. One-tailed hypothesis tests were used for TES and p-uniform's publication bias test, because only evidence in one direction for these methods is indicative of publication bias. For each simulated meta-analysis, we recorded the proportion of data sets for which the statistical power of a publication bias test was larger than 0.8. Meta-analyses were simulated 10,000 times for all included data sets. True effect size μ was fixed to zero for generating data, because this enabled simulating data using the same effect size measure as in the data sets. Selected values for publication bias (pub) were 0, 0.25, 0.5, 0.75, 0.85, 0.95, and 1 where pub equal to 0 indicates no publication bias and 1 extreme publication bias. This Monte-Carlo simulation study was programmed in R 3.5.3 (R Core Team, 2019) and the packages "metafor" (Viechtbauer, 2010), "puniform" (van Aert, 2019), and "parallel" (R Core Team, 2019) were used. R code for this Monte-Carlo simulation study is available at https://osf.io/pg7sj.

Estimating effect size corrected for publication bias. Five different methods were included to estimate the effect size: traditional meta-analysis, trim and fill, PET-PEESE, p-uniform, and the selection model approach proposed by Vevea and Hedges (1995). Traditional meta-analysis was included, because it is the analysis that is conducted in every meta-analysis. Either a fixed-effect (FE) or random-effects (RE) model was selected depending on the statistical model used in the meta-analysis. These publication bias methods were selected, because they were either often applied in meta-analyses (trim and fill) or outperformed other methods (PET-PEESE, p-uniform, and the selection model approach; McShane et al., 2016; Stanley & Doucouliagos, 2014; van Assen et al., 2015). P-curve (Simonsohn et al., 2014) was not included in the present study because the methodology underlying p-curve is the same as p-uniform, and p-uniform has the advantage that it can also test for publication bias and estimate a 95% confidence interval (CI; see https://osf.io/b9t7v/ for an elaborate overview of existing methods to correct effect size for publication bias).

Average effect size estimates of traditional meta-analysis, trim and fill, PET-PEESE, p-uniform, and the selection model approach were computed and transformed to a common effect size measure (i.e., Cohen's d) before interpreting them. Data sets that used log relative risks as effect size measure were conducted based on log odds ratios and these average effect size estimates were transformed to Cohen's d values. If there was not enough information to transform Hedges' g to Cohen's d, Hedges' g was used in the analyses. Effect sizes were computed using the formulas described in Borenstein (2009).

We assessed the severity of publication bias by computing difference scores in effect size estimates between traditional meta-analysis and each publication bias method (i.e., trim and fill, PET-PEESE, p-uniform, and the selection model approach). That is, we subtracted the effect size estimate of traditional meta-analysis from the method's effect size estimate. A difference score of zero reflects that the estimates of traditional meta-analysis and the publication bias method were the same, whereas a positive or negative difference score indicates that the estimates were different. Subsequently, the mean and standard deviation (SD) of these difference scores were computed for the three methods.

All analyses were conducted using R version 3.5.3 (R Core Team, 2019). The "metafor" package (Viechtbauer, 2010) was used for conducting fixed-effect or random-effects meta-analysis, trim and fill, rank-correlation test, and Egger's regression test. The "puniform" package (van Aert, 2019) was used for applying the p-uniform method using the default estimator based on the Irwin-Hall distribution. In line with the recommendation by Stanley (2017), α=0.1 was used for the right-tailed test whether the intercept of a PET analysis was different from zero, and therefore whether the results of PET or PEESE had to be interpreted. The selection model approach as proposed by (Vevea & Hedges, 1995) and implemented in the "weightr" package (Coburn & Vevea, 2019) was applied to all data sets. Data and R code of

the analyses are available at https://osf.io/afnvr/ and https://osf.io/taq5f/?.

Results

Description of Meta-Analyses investigated

A flowchart illustrating the procedure of selecting meta-analyses and data sets is presented in Figure 1. The literature search resulted in 7,647 hits including duplicates, the screening process reduced this number to 502 meta-analyses, of which 89 dealt with the efficacy of psychotherapeutic interventions for PTSD and were included (see Appendix A and https://osf.io/pkzx8/). Of these 89 meta-analyses, four could not be located as they were unpublished dissertations and the authors did not reply to our requests.[2] One meta-analysis was excluded because it used a network meta-analysis approach (Gerger et al., 2014) and the included publication bias methods cannot be applied to this type of data. A multi-site study (Morrissey et al., 2015) was excluded, because meta-analysis methods were used to combine the results from the different sites. Of the remaining 83 meta-analyses, we contacted 36 authors (43.4%) because the effect size data was not fully reported in their paper and obtained data from six authors (16.7%).

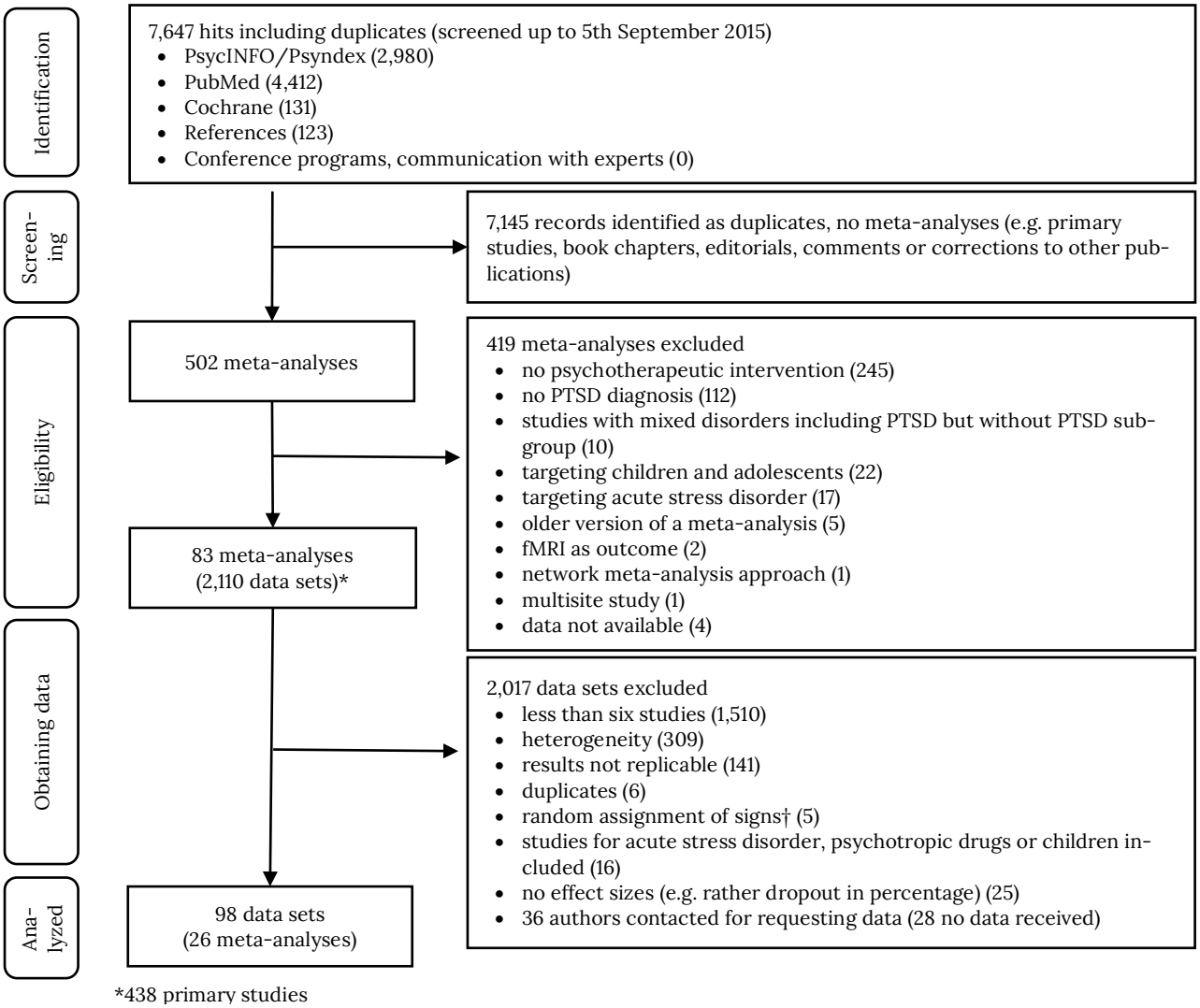

Figure 1. Flow chart: Identification and selection of meta-analyses and data sets. Note. † positive and negative signs were randomly assigned to each effect in the meta-analysis

Meta-Psychology, 2020, vol 4, MP.2018.884,
https://doi.org/10.15626/MP.2018.884
Article type: Original Article
Published under the CC-BY4.0 license

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Daniël Lakens
Reviewed by: K. Coburn, F. D. Schönbrodt
Analysis reproduced by: Erin M. Buchanan
All supplementary files can be accessed at the OSF project page: https://doi.org/10.17605/OSF.IO/MY8S6

Our analysis of the 83 meta-analyses first examined whether they discussed the problem of publication bias. Fifty-eight meta-analyses (69.9%) mentioned publication bias, whereas 25 (30.1%) did not mention it at all. In 35 meta-analyses (42.2%), it was specified that the search strategies included unpublished studies, and 20 (24.1%) indeed found and included unpublished studies. However, in 46 meta-analyses (55.4%) unpublished studies were explicitly regarded as unsuitable for inclusion, and two meta-analyses (2.4%) did not specify their search and inclusion criteria with respect to unpublished studies.

Forty-seven meta-analyses (56.6%) statistically assessed publication bias, whereas 36 (43.4%) did not. Five meta-analyses (6.0%) included the rank-correlation test, six (7.2%) Egger's regression test, and nine (10.8%) the trim and fill procedure. TES, PET-PEESE and p-uniform were not applied in any of the meta-analyses. A funnel plot (Light & Pillemer, 1984) was presented in 26 meta-analyses (31.3%) and failsafe N (Rosenthal, 1979) was computed in 26 meta-analyses (31.3%). These results indicate that a large number of meta-analyses did not assess publication bias or only applied a selection of publication bias methods. PET-PEESE and p-uniform have been developed more recently and therefore we did not expect them to be regularly applied.

The 83 meta-analyses included a total number of 2,110 data sets, of which 98 (4.6%) data sets from 26 meta-analyses fulfilled all inclusion criteria and were eligible for publication bias assessment (see flowchart in Figure 1). Figure 2 is a histogram of the number of effect sizes per data sets before data sets were excluded due to less than six studies and heterogeneous true effect size. The results show that the majority of data sets contained less than six effect sizes, and that only a small number of data sets included more than 15 effect sizes.

Many data sets were excluded because there were less than six studies (1,510 data sets), and due to heterogeneity in true effect size (309 data sets). All meta-analyses of which data sets were included in our study are marked with an asterisk in the list of references.

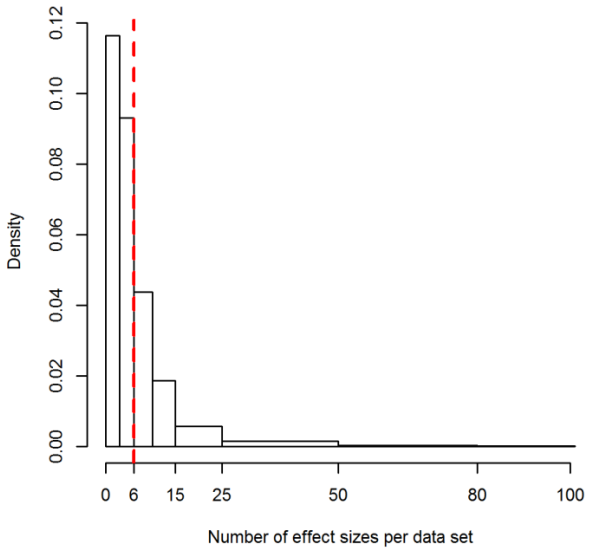

Figure 2. Histogram of the number of primary studies' effect sizes included in data sets. The vertical red dashed line denotes the cut-off that was used for assessing publication bias in a meta-analysis.

Characteristics of included data sets

Thirty-nine (39.8%) data sets reported Hedges' g as effect size measure, 29 (29.6%) Cohen's d, 3 (3.1%) a standardized mean difference, 7 (7.1%) a raw mean difference, 16 (16.3%) risk ratio, 2 (2.0%) log odds ratio, and 2 (2.0%) data sets Glass' delta.

The median number of effect sizes in a data set was 7 (first quartile 7, third quartile 10). Since publication bias tests have low statistical power if the number of effect sizes is small in a meta-analysis (Begg & Mazumdar, 1994; Sterne et al., 2000; van Assen et al., 2015), the characteristics of many of the data sets are not well-suited for methods to detect publication bias. Additionally, p-uniform cannot be applied if there are no statistically significant effect sizes in a meta-analysis, because a requirement is that at least one study in a meta-analysis is statistically significant. The median number of statistically significant effect sizes in the data sets was 3 (34.3%; first quartile 1 (13%), third quartile 6 (80.4%)), and 77 data sets (78.6%) included at least one significant effect size (see Appendix A, which also reports the number of studies included in each data set).

Consequently, conditions were also not well-suited for p-uniform in particular, since this method uses only the statistically significant effect sizes. The median I^2-statistic was 0% (first quartile 0%, third quartile 28.7%).

Publication Bias Test

Before applying the publication bias tests to the data sets, we conducted a Monte-Carlo simulation study to examine whether statistical power of the tests is large enough (> 0.8) to warrant applying these tests. Type-I error rate and statistical power of the rank-correlation test (open circles), Egger's test (triangles), TES (diamonds), and p-uniform's publication bias test (solid circles) as a function of publication bias (pub) are shown in Figure 3. The results in the figure were obtained by averaging over the 98 data sets and the 10,000 replications in the Monte-Carlo simulation study. Type-I error rate of all publication bias tests was smaller than α =.05 implying that the tests were conservative. These results indicate that statistical power of all methods was not above 0.5 for pub < 0.95. Statistical power of only the TES was larger than 0.8 in case of extreme publication bias (pub = 1).

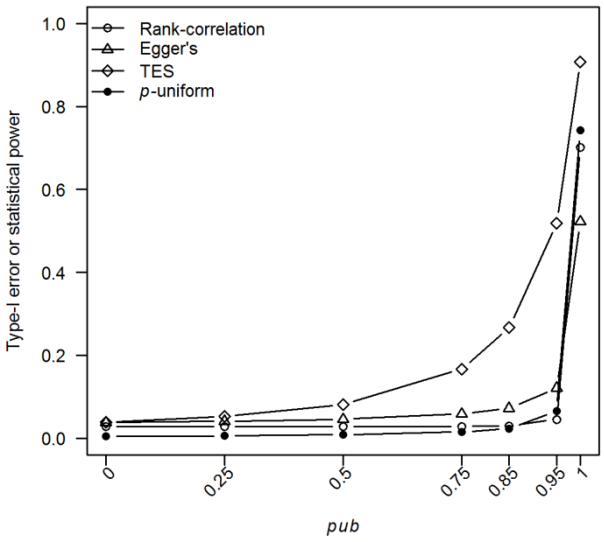

Figure 3. Type-I error rate and statistical power obtained with the Monte-Carlo simulation study of the rank-correlation test (open circles), Egger's test (triangles), test of excess significance (TES; diamonds), and p-uniform's publication bias test (solid circles)

We also studied in the simulations whether for each data set the statistical power of a publication bias test was larger than 0.8. This enabled us to select the data sets where publication bias tests would be reasonable powered to detect publication bias if it was present. Statistical power of none of the methods was larger than 0.8 for any data set if pub < 0.95 (results are available at https://osf.io/6bnc5/ for the rank-correlation test, https://osf.io/ufdps/ for Egger's test, https://osf.io/5yehp/ for the TES, and https://osf.io/feux3/ for p-uniform). It is highly unlikely that publication bias is this extreme in the included data sets, because many data sets contained statistically nonsignificant effect sizes (median percentage of nonsignificant effect sizes in a data set 65.7%). The publication bias tests would be most likely severely underpowered when applied to the published meta-analyses on PTSD, and it follows from these results that the tests should not be applied here. Therefore, we only report the results of applying the publication bias tests to the data sets as supplement in the online repository (https://osf.io/49cke/) for completeness.

Effect Size Corrected for Publication Bias

The data set with ID 77 (from the meta-analysis by Kehle-Forbes et al., 2013) was excluded for estimating effect sizes corrected for publication bias because not enough information was available to transform the log relative risks to Cohen's d. Hedges' g effect sizes could not be transformed into Cohen's d for 12 data sets and Hedges' g was used instead (see Appendix A). Descriptive results of the effect size estimates of traditional meta-analysis, trim and fill, PET-PEESE, p-uniform, and the selection model approach are presented in Table 1. P-uniform could only be applied to data sets with at least one statistically significant result (77 data sets), and the selection model approach did not converge for two data sets. Results showed that especially estimates of PET-PEESE were closer to zero than traditional meta-analysis and that the standard deviation of the estimates of PET-PEESE and p-uniform was larger than traditional meta-analysis, trim and fill, and the selection model approach. See Appendix A for the results of the effect size estimates corrected for publication bias per data set.

Table 1.
Descriptive results of data sets analyzed with meta-analysis (fixed-effect or random-effects model depending on the model that was used in the original meta-analysis), trim and fill, PET-PEESE, p-uniform, and the selection model approach.

	Mean, median	[min.; max.], (SD) of estimates
Table format	0.603, 0.532	[0.015;1.85], (0.447)
Trim and fill	0.574, 0.467	[-0.047;1.789], (0.411)
PET-PEESE	0.219, 0.203	[-1.656;3.075], (0.696)
p-uniform	0.556, 0.693	[-6.681;2.158], (1.385)
Sel. model	0.603, 0.536	[-0.061;1.828], (0.439)

Note. min. is the minimum value, max. is the maximum value, and SD is the standard deviation.

The mean of the difference in effect size estimate between PET-PEESE and the meta-analytic estimate was -0.101 (SD = 0.872). However, the median of the difference in effect size estimate was close to zero (Mdn = -0.002), suggesting that the estimates of PET-PEESE and traditional meta-analysis were close. The mean of the difference between the estimates of trim and fill and traditional meta-analysis (-0.009, Mdn = 0, SD = 0.104) and the selection model approach and traditional meta-analysis was negligible (0.026, Mdn = 0.026, SD = 0.145).

Analyses for data sets including significant effect sizes. P-uniform was applied to a subset of 77 data sets (see Appendix A), because this method requires that at least one study is statistically significant. The mean of the difference in effect size estimate of p-uniform and traditional meta-analysis was -0.174 (Mdn = 0.04, SD = 1.273). The large standard deviation is caused by situations in which an extreme effect size was estimated because a primary study's effect size was only marginally significant (i.e., p-value just below .05). In order to counteract these extreme effect size estimates, we set p-uniform's effect size estimate to zero when the average of the statistically significant p-values was larger than half the α-level.[3] This is in line with the recommendation by van Aert et al. (2016). Setting this effect size to zero resulted in a mean of the difference in effect size estimate between p-uniform and traditional meta-analysis of -0.019 (Mdn = 0.04, SD = 0.364). The change in difference in effect size estimate was caused by setting the effect size estimates of p-uniform in seven data sets to zero, in which p-uniform originally substantially corrected for publication bias. The mean of the difference scores between PET-PEESE and traditional meta-analysis when computed based on this subset of 77 data sets was -0.129 (Mdn = -0.011, SD = 0.968), for trim and fill the mean of the difference scores was -0.014 (Mdn = 0, SD = 0.105), and for the selection model approach the mean of the difference scores was 0.028 (Mdn = 0.024, SD = 0.155).

Explaining estimates of p-uniform, the selection model approach, and PET-PEESE. We illustrate deficiencies of p-uniform, the selection model approach, and PET-PEESE by discussing the results of two exemplary data sets. Estimates of p-uniform can be imprecise (i.e., with a wide CI) if they are based on a small number of effect sizes in combination with p-values of these effect sizes close to the α-level. In 29 out of 77 data sets p-uniform's estimate was based on at most three studies. For instance, the estimated average log relative risk of random-effect meta-analysis of the data set from Bisson et al. (2013, ID=20) was -0.177, 95% CI [-0.499, 0.145] and p-uniform's estimate was based on a single study and equaled -0.504, 95% CI [-3.809, 8.174]. The effect size estimate of p-uniform, as for any other method, is more precise the larger the number of effect sizes in a data set or the larger the primary study's sample sizes.

The selection model approach also suffers from a small number of statistically significant effect sizes. The computed weights for the intervals of the method's selection model are imprecisely estimated if only a small number of effect sizes are within an interval. In an extreme situation where no effect sizes are observed in an interval of the selection model, the implementation of the selection model approach by Vevea and Hedges (1995) in the R package "weightr" assigns a weight of 0.01 to this interval. Bias in effect size estimation increases the more this weight deviates from its true value.

PET-PEESE also did not result in reasonable effect size estimates in each of the data sets, and

especially not if the standard errors of the primary studies were highly similar (i.e., were based on similar sample sizes). Figure 4 shows the funnel plot based on the data set from Bisson et al. (2007) comparing TF-CBT versus wait list and active controls (ID=14; left panel) with the filled circles being the 15 observed effect sizes. The studies' standard errors diverged from each other, which makes it possible to fit a regression line through the observed effect sizes in the data set (dashed black line). PET-PEESE's effect size estimate was -0.027 (95% CI [-0.663, 0.609]) denoted by the asterisk in Figure 4), which was closer to zero than traditional meta-analysis (0.260, 95% CI [-0.057, 0.578]) but had a wider CI. The data set from Diehle et al. (2014) comparing two different treatments of TF-CBT (ID=44) is presented in the right panel of Figure 4. PET-PEESE was hindered by the highly similar studies' standard error, which ranged from 0.227 to 0.478. Hence, the effect size estimate of PET-PEESE (0.44, 95% CI [-1.079, -1.958]) was unrealistically larger than the estimate traditional meta-analysis (-0.153, 95% CI [-0.084, 0.39]), and its CI was wider.

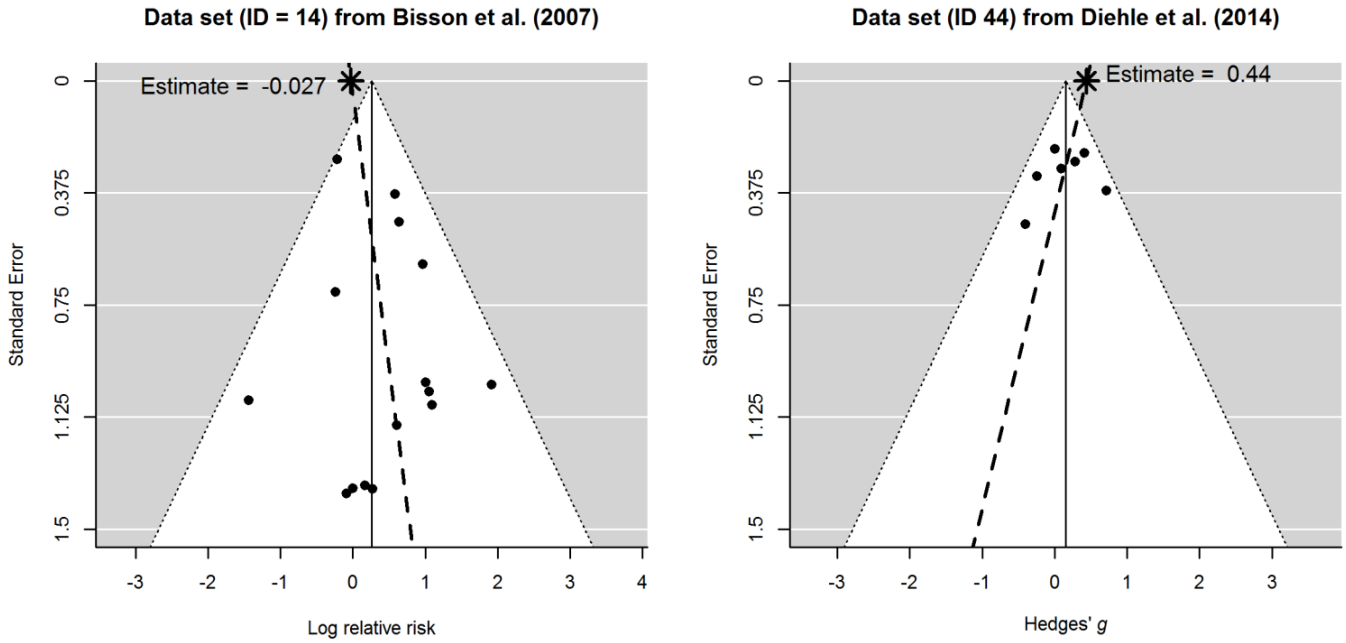

Figure 4. Funnel plots of the data sets from Bisson et al. (2013) (ID=14; left panel) and Diehle et al. (2014) (ID=44; right panel). Filled circles are the observed effect sizes in a meta-analysis, the dashed black line is the fitted regression line through the observed effect sizes, the asterisks indicate the estimate of PET-PEESE.

Discussion

Publication bias is widespread in the psychology research literature (Bakker et al., 2012; Fanelli, 2012; Sterling et al., 1995) resulting in overestimated effect sizes in primary studies and meta-analyses (Kraemer, Gardner, Brooks, & Yesavage, 1998; Lane & Dunlap, 1978). Guidelines such as the MARS (American Psychological Association, 2010) and PRISMA (Moher, Liberati, Tetzlaff, & Altman, 2009) recommend to routinely correct for publication bias in any meta-analysis. Others recommend to re-analyze published meta-analyses to study the extent of publication bias in whole fields of research (Ioannidis, 2009; Ioannidis & Trikalinos, 2007a; van Assen et al., 2015) by using multiple publication bias methods (Coburn & Vevea, 2015; Kepes et al., 2012). However, the question is whether routinely assessing publication bias is indeed a good recommendation, because researchers may end up in applying publication bias methods in situations where these do not have appropriate statistical properties, potentially leading to drawing faulty conclusions. We tried to answer

this question by re-analyzing a large number of meta-analyses published on the efficacy of psychotherapeutic treatment for PTSD.

We re-analyzed 98 data sets from 26 meta-analyses studying a wide variety of psychotherapeutic treatments for PTSD. We had to exclude a large portion of data sets (95.4%) mainly due to heterogeneity in true effect size and data sets containing less than six primary studies. These exclusion criteria were necessary, because publication bias methods do not perform well in case of heterogeneity in true effect size (Ioannidis, 2005) and a small number of primary studies yields low power of publication bias methods and imprecise effect size estimation (Sterne et al., 2000).

The included data sets were characterized by including a small number of primary studies (median 7 studies) resulting in challenging conditions for any publication bias method. Before applying publication bias tests, we studied whether these tests would have sufficient statistical power (>0.8). We conducted a Monte-Carlo simulation study in which data were generated in a way to stay as close as possible to the included data sets. The statistical power of the publication bias tests was only larger than 0.8 in case of extreme publication bias (i.e., nonsignificant effect sizes having a probability of 0.05 or smaller to be included in a meta-analysis). Hence, we concluded that it was not warranted to apply publication bias tests. Of note is that the median percentage of nonsignificant effect sizes in a data set was 65.7% suggesting that extreme publication bias was absent.

Publication bias methods that correct the effect size for bias are also affected by a small number of primary studies, because the effect size estimates become then imprecise (i.e., a wide CI). However, comparing estimates of these methods with those of traditional meta-analysis that does not correct for publication bias still provides insights about the severity of publication bias. This analysis revealed no evidence for severe overestimation caused by publication bias as the corrected estimates were close to those of traditional meta-analysis.

Our results imply that following up on the guidelines to assess publication bias in any meta-analysis is far from straightforward in practice. Many data sets in our study where too heterogeneous for publication bias analyses. Moreover, even after the exclusion of data sets with less than six studies, statistical power of publication bias tests for each data set was low if extreme publication bias was absent and CIs of methods that provided estimates corrected for publication bias were wide. These results even call for revising the recommendation by Sterne et al. (2011) to apply publication bias tests only to meta-analyses with more than 10 studies. Our results are also corroborated by a recent study of Renkewitz and Keiner (2019) who concluded based on a simulation study that publication bias could only be reliably detected with at least 30 studies. However, a caveat here is that these recommendations heavily depend on the severity of publication bias that is assumed to be present in a meta-analysis. Hence, most important is that researchers are aware of the fact that publication bias tests suffer from low statistical power and that a nonsignificant publication bias test does not imply that publication bias is absent.

Recommendations

We consider it important to give practical advice to researchers. We recommend researchers to follow the MARS guidelines, apply publication bias tests, and report effect size estimates corrected for publication bias. However, a well-informed choice has to be made to select the publication bias methods with the best statistical properties as no method outperforms all other methods in all conditions (Carter et al., 2019; Renkewitz & Keiner, 2019). Carter and colleagues (2019) conclude that it has not been investigated yet whether the application of publication bias methods is warranted in real data in psychology, and that this ultimately is an empirical question which should be the focus of future research. Routinely applying publication bias methods without paying attention to their statistical properties for the characteristics of the respective meta-analysis cannot be recommended. Hence, researchers need to consider the characteristics of the data sets and check the properties of publication bias methods for these data sets before actually applying these methods. Such a "method performance check" has also been recommended by Carter et al. (2019) for methods to correct effect size for publication bias and can be conducted by their meta explorer web application (http://www.shinyapps.org/apps/metaExplorer/) or simulation studies. A complicating factor,

however, is that a method performance check requires information about the true effect size, true heterogeneity in true effect size, and the extent of publication bias that is not available. Hence, researchers are advised to use multiple levels for these parameters in a method performance check as a sensitivity analyses.

As there is no single publication bias method that outperforms all other methods and selecting a method depends on unknown parameters, we recommend to apply multiple publication bias methods that show acceptable performance in a method performance check. A so-called triangulation (Kepes et al., 2012; Coburn & Vevea, 2015) following a methods performance check, rather than applying only one publication bias method, will yield more insight into the presence and severity of publication bias, because each method uses its own approach to examine publication bias. Researchers should refrain from testing for publication bias if a method performance check by means of a power analysis reveals that publication bias is unlikely to be detected in their meta-analysis. Applying methods to correct effect size for publication bias is still useful in case of a small number of studies in a meta-analysis, because estimates corrected for publication bias can be compared to the uncorrected estimate to assess the severity of publication bias.

We consider it important to emphasize that the reporting of publication bias methods should be independent of their results. The analysis procedure of the meta-analysis as well as the publication bias tests is preferred to be preregistered in a pre-analysis plan before the analyses are actually conducted. Moreover, conflicting results of publication bias methods are an interesting and important finding on its own that has to be discussed in the paper.

Limitations

Heterogeneous data sets had to be excluded, because assessing publication bias with the included methods is only accurate when based on meta-analyses with no or small heterogeneity in true effect size (Ioannidis & Trikalinos, 2007a; Terrin et al., 2003). For that reason, data sets were excluded from the analyses if the I^2-statistic was larger than 50%, but the I^2-statistic is generally imprecise and especially if the number of effect sizes in a meta-analysis is small (Ioannidis, Patsopoulos, & Evangelou, 2007). This is also reflected in the wide confidence intervals around the I^2-statistics of the included data sets in the analyses (see Appendix A). Moreover, there is also an effect of publication bias on the I^2-statistic which has been shown to be large, complex and non-linear, such that publication bias can both dramatically decrease and increase the I^2-statistic (Augusteijn, van Aert, & van Assen, 2019). Therefore, a consequence of using a selection criterion based on the I^2-statistic in the current study is that this may have led to the inclusion of data sets with heterogeneity in true effect size, which may, in turn, also have biased the results of the publication bias methods because these methods do not perform well under substantial heterogeneity (Ioannidis, 2005; Terrin et al., 2003; van Assen et al., 2015).

Data sets affected by publication bias may also have been excluded by limiting ourselves to homogeneous data sets. Imagine a data set consisting of multiple statistically significant effect sizes because of publication bias and one nonsignificant effect size that is not influenced by publication bias. The inclusion of this nonsignificant effect size likely causes the I^2-statistic to be larger than 50% while the true effect size in fact may be homogeneous. Hence, publication bias may also have resulted in the exclusion of homogeneous data sets. Another limitation is that questionable research practices, known as p-hacking (i.e., all behaviors researchers can use to obtain the desired results; Simmons, Nelson, & Simonsohn, 2011), may have further biased the results of the publication bias methods as well as the traditional meta-analysis (van Aert et al., 2016).

Of note is also that the data sets in the current investigation often contained multiple statistically nonsignificant effect sizes when an active treatment was compared to a passive or active control group, which is not expected in case of extreme publication

bias. Especially comparisons between two active treatments resulted in very few significant differences in efficacy. These meta-analyses with nonsignificant comparative effects might also be affected by publication bias. For example, when a new treatment is found to be as efficacious as an established one, this might be newsworthy and have a larger chance to get published than a finding demonstrating the well-known superiority of the state-of-the-art treatment. This implies that publication bias lead to the publication of statistically nonsignificant rather than significant effects. Publication bias will not be detected by any of the methods in such a situation in this study.

Conclusion

Routinely assessing publication bias in any meta-analysis is recommended by guidelines such as MARS and PRISMA. We have shown, however, that the characteristics of meta-analyses in research on PTSD treatments are generally unfavorable for publication bias methods. That is, heterogeneity and small numbers of studies in meta-analyses result in low statistical power and imprecise corrected estimates. Of note is that interpreting results from small data sets cautiously accounts in general for meta-analyses. The characteristics of the meta-analyses in our study on PTSD treatments are deemed to be typical for psychotherapy research, and potentially for other areas of clinical psychology, as well.

The development of new publication bias methods and the improvement of existing methods is necessary that allow the true effect size to be heterogeneous and perform well in case of a small number of effect sizes in a meta-analysis. Promising developments are p-uniform being extended to enable accurate effect size estimation in the presence of heterogeneity in true effect size (van Aert & van Assen, 2020). Other promising developments are Bayesian methods to correct for publication bias (Du, Liu, & Wang, 2017; Guan & Vandekerckhove, 2016) and the increased attention for selection model approaches (Citkowicz & Vevea, 2017; McShane et al., 2016).

We hope that our work creates awareness for the limitations of publication bias methods and recommend researchers to apply and report multiple publication bias methods that have shown good statistical properties for the meta-analysis under study.

Author Contact

Helen Niemeyer
Freie Universität Berlin
Department of Clinical Psychological Intervention
Schwendenerstr. 27
14195 Berlin
Germany

Phone 0049/3083854798
helen.niemeyer@fu-berlin.de

Conflict of Interest and Funding

All authors declare no conflict of interest. The preparation of this article was supported by Grant 406-13-050 from the Netherlands Organization for Scientific Research (RvA).

Author Contributions

HN and RvA designed the study. HN and SSch developed the search strategy and SSch coordinated the literature search. SSch, HN and DU served as independent raters in the process of the selection of the meta-analyses. HN and SSch conducted the data collection, coding and data management. SSch contacted the authors of all primary studies for which data were missing. DU developed the treatment coding scheme, and HN, DU and SSch categorized the treatment and control conditions. RvA performed all statistical analyses. HN and RvA drafted the manuscript. CK and OSH supervised the whole conduction of the meta-meta-analysis and revised the manuscript.

Open Science Practices

This article earned the Open Data and the Open Materials badge for making the data and materials openly available. It has been verified that the analysis reproduced the results presented in the article. It should be noted that the coding of the literature has not been verified; only the final analysis. The

entire editorial process, including the open reviews, are published in the online supplement.

Acknowledgements

The authors would like to thank Helen-Rose and Sinclair Cleveland, Andrea Ertle, Manuel Heinrich, Marcel van Assen, Josephine Wapsa and Jelte Wicherts who helped in proof reading the article.

References

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2010). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. Journal of Management, 37(1), 5-38. doi:10.1177/0149206310377113

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. Journal of Organizational Behavior, 32(8), 1033-1043. doi:10.1002/job.719

American Psychiatric Association. (1987). Diagnostic and statistical manual of mental disorders (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, DC: Author.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). Publication manual of the American Psychological Association. Washington, DC: Author.

Augusteijn, H. E. M., van Aert, R. C. M., & van Assen, M. A. L. M. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. Psychological Methods, 24(1), 116-134. doi:10.1037/met0000197

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. Perspectives On Psychological Science, 7(6), 543-554. doi:10.1177/1745691612459060

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: the antagonist of meta-analytic reviews and effective policymaking. Educational Evaluation and Policy Analysis, 34(3), 259-277. doi:10.3102/0162373712446144

Banks, G. C., Kepes, S., & McDaniel, M. A. (2012). Publication bias: a call for improved meta-analytic practice in the organizational sciences. International Journal of Selection and Assessment, 20(2), 182-197. doi:10.1111/j.1468-2389.2012.00591.x

Becker, B. J. (2005). Failsafe N or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 111-125). Chichester, England: Wiley.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. Biometrics, 50, 1088-1101.

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(5), 746-758. doi:10.1016/j.cpr.2007.10.005

Berlin, J. A., & Ghersi, D. (2005). Preventing publication bias: Registries and prospective meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 35-48). Chichester, England: Wiley.

Bisson, J. I., Roberts, N. P., Andrew, M., Cooper, R., & Lewis, C. (2013). Psychological therapies for chronic post-traumatic stress disorder (PTSD) in adults. Cochrane Database of Systematic Reviews, 12. doi:10.1002/14651858.CD003388.pub4

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (pp. 221-236). New York, NY: Russell Sage Foundation.

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

Bornstein, H. A. (2004). A meta-analysis of group treatments for post-traumatic stress disorder: How treatment modality affects symptoms. (64), ProQuest Information & Learning, US. Retrieved from

http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2004-99008-373&site=ehost-live Available from EBSCOhost psyh database.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. Advances in Methods and Practices in Psychological Science, 2, 1 - 24. Retrieved from osf.io/preprints/psyarxiv/9h3nu

Chard, K. M. (1995). A meta-analysis of posttraumatic stress disorder treatment outcome studies of sexually victimized women. (55), ProQuest Information & Learning, US. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1995-95007-211&site=ehost-live Available from EBSCOhost psyh database.

Citkowicz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. Psychological Methods, 22(1), 28-41. doi:doi:10.1037/met0000119

Coburn, K. M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. Psychological Methods, 20(3), 310-330. doi:10.1037/met0000047

Coburn, K. M., & Vevea, J. L. (2019). weightr: Estimating Weight-Function Models for Publication Bias. R package version 2.0.1. doi: https://CRAN.R-project.org/package=weightr

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 11-33). Chichester, England: Wiley.

Driessen, E., Hollon, S. D., Bockting, C. L. H., & Cuijpers, P. (2017). Does Publication Bias Inflate the Apparent Efficacy of Psychological Treatment for Major Depressive Disorder? A Systematic Review and Meta-Analysis of US National Institutes of Health-Funded Trials. PLOS ONE, 10(9), e0137864. doi:doi:10.1371/journal.pone.0137864

Du, H., Liu, F., & Wang, L. (2017). A Bayesian "fill-In" method for correcting for publication bias in meta-analysis. Psychological Methods, 22(4), 799-817. doi:10.1037/met0000164

Duval, S., & Tweedie, R. (2000a). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. Journal of the American Statistical Association, 95(449), 89-98.

Duval, S., & Tweedie, R. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics, 56(2), 455-463. doi:10.1111/j.0006-341X.2000.00455.x

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. Brithish Medical Journal, 315(7109), 629-634. doi:10.1136/bmj.315.7109.629

Ehlers, A., Bisson, J., Clark, D. M., Creamer, M., Pilling, S., Richards, D., . . . Yule, W. (2010). Do all psychological treatments really work the same in posttraumatic stress disorder? Clinical Psychology Review, 30(2), 269-276.

Ehlers, A., Clark, D. M., Hackmann, A., McManus, F., & Fennell, M. (2005). Cognitive therapy for post-traumatic stress disorder: development and evaluation. Behaviour research and therapy, 43(4), 413-431.

Ellis, P. D. (2010). The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. New York: Cambridge University Press.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. Scientometrics, 90(3), 891-904. doi:10.1007/s11192-011-0494-7

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. Psychological Methods, 17(1), 120-128. doi:10.1037/a0024445

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. British Journal of Mathematical and Statistical Psychology, 63(3), 665-694. doi:10.1348/000711010X502733

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-382. doi:10.1037/h0031619

Foa, E. B., & Rothbaum, B. O. (1998). Treating the trauma of rape: Cognitive-behavioral therapy for PTSD. New York, NY: Guilford Press.

Forbes, D., Creamer, M., Bisson, J. I., Cohen, J. A., Crow, B. E., Foa, E. B., . . . Ursano, R. J. (2010). A

guide to guidelines for the treatment of PTSD and related conditions. Journal of traumatic stress, 23(5), 537-552.

Forbes, D., Creamer, M., Phelps, A., Bryant, R., McFarlane, A., Devilly, G. J., . . . Newton, S. (2007). Australian guidelines for the treatment of adults with acute stress disorder and post-traumatic stress disorder. Aust N Z J Psychiatry, 41(8), 637-648. doi:10.1080/00048670701449161

Francis, G. (2013). Replication, statistical consistency, and publication bias. Journal of Mathematical Psychology, 57(5), 153-169. doi:10.1016/j.jmp.2013.02.003

Gerger, H., Munder, T., Gemperli, A., Nüesch, E., Trelle, S., Jüni, P., & Barth, J. (2014). Integrating fragmented evidence by network meta-analysis: relative effectiveness of psychological interventions for adults with post-traumatic stress disorder. Psychol Med, 44(15), 3151-3164. doi:10.1017/S0033291714000853

Gilbody, S., & Song, F. (2000). Publication bias and the integrity of psychiatry research. Psychological Medicine, 30, 253-258. doi:10.1017/S0033291700001732

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. Psychonomic Bulletin & Review, 23(1), 74-86. doi:doi:10.3758/s13423-015-0868-6

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. PLoS Biol, 13(3), e1002106. doi:10.1371/journal.pbio.1002106

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 145-174). Chichester, UK: Wiley.

Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. Statistics in Medicine, 21(11), 1539-1558. doi:10.1002/sim.1186

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). Applied statistics for the behavioral sciences. Boston, Mass.: Houghton Mifflin Company.

Hopewell, S., Clarke, M., & Mallett, S. (2005). Grey literature and systematic reviews. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 49-72). Chichester, England: Wiley.

Ioannidis, J. P. (2005). Differentiating biases from genuine heterogeneity: distinguishing artifactual from substantive effects. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: prevention, assessment and adjustments (pp. 287-302). Sussex, England: Wiley.

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. Epidemiology, 19(5), 640-648. doi:10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. (2009). Integration of evidence from multiple meta-analyses: A primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. Canadian Medical Association Journal, 181(8), 488-493. doi:10.1503/cmaj.081086

Ioannidis, J. P., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. British Medical Journal, 335(7626), 914-916. doi:10.1136/bmj.39343.408449.80

Ioannidis, J. P., & Trikalinos, T. A. (2007a). The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. Canadian Medical Association Journal, 176(8), 1091-1096. doi:10.1503/cmaj.060410

Ioannidis, J. P., & Trikalinos, T. A. (2007b). An exploratory test for an excess of significant findings. Clinical Trials, 4(3), 245-253. doi:10.1177/1740774507079441

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. Statistical Science, 3, 109-135.

Jaycox, L. H., & Foa, E. B. (1999). Cost-effectiveness issues in the treatment of post-traumatic stress disorder. In N. E. Miller & M. K. M (Eds.), Cost-effectiveness of psychotherapy: A guide for practitioners, researchers, and policymakers. New York, NY: Oxford University Press.

Karen, R. M. (1990). Shame and guilt as the treatment focus in Post-Traumatic Stress Disorder: A meta-analysis. (51), ProQuest Information & Learning, US. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1991-51715-001&site=ehost-live Available from EBSCOhost psyh database.

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. Organizational Research Methods, 15(4), 624-662.

Kessler, R. C., Petukhova, M., Sampson, N. A., Zaslavsky, A. M., & Wittchen, H. U. (2012). Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. International journal of methods in psychiatric research, 21(3), 169-184.

Kraemer, H. C., Gardner, C., Brooks, J., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. Psychological Methods, 3(1), 23-31. doi:doi:10.1037/1082-989X.3.1.23

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. British Journal of Mathematical and Statistical Psychology, 31(2), 107-112.

Light, R., & Pillemer, D. (1984). Summing up: The science of research reviewing. In. Cambridge, MA: Harvard University Press.

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

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of scale analysis and factor analysis. Psychological Bulletin, 45(6), 507-529. doi:10.1037/h0055827

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. Statistics in Medicine, 20(4), 641-654.

Maljanen, T., Knekt, P., Lindfors, O., Virtala, E., Tillman, P., Härkänen, T., & Helsinki Psychotherapy Study Group. (2016). The cost-effectiveness of short-term and long-term psychotherapy in the treatment of depressive and anxiety disorders during a 5-year follow up. Journal of Affective Disorders, 190, 254-263. doi:10.1016/j.jad.2015.09.065

Margraf, J. (2009). Kosten und Nutzen der Psychotherapie. Heidelberg: Springer Medizin.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. Perspectives On Psychological Science, 11(5), 730-749. doi:doi:10.1177/1745691616662243

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Annals of internal medicine, 151(4), 264-269.

Mokken, R. J. (1971). A theory and procedure of scale analysis: With applications in political research (Vol. 1): Walter de Gruyter.

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. BMC Medical Research Methodology, 9(2). doi:10.1186/1471-2288-9-2

National Collaborating Centre for Mental Health. (2005). Post-traumatic stress disorder: The management of PTSD in adults and children in primary and secondary care. (NICE Clinical Guidelines, No. 26). In. Leicester, UK: Gaskell.

Niemeyer, H., Musch, J., & Pietrowsky, R. (2012). Publication bias in meta-analyses of the efficacy of psychotherapeutic interventions for schizophrenia. Schizophrenia research, 138(2), 103-112.

Niemeyer, H., Musch, J., & Pietrowsky, R. (2013). Publication bias in meta-analyses of the efficacy of psychotherapeutic interventions for depression. Journal of consulting and clinical psychology, 81(1), 58-74.

Niemeyer, H., Pieper, A., Uelsmann, D., Schulte-Herbrüggen, O., & Knaevelsrud, C. (2017). Evidence based psychotherapy for complex post-traumatic stress disorder (PTSD), and PTSD following complex traumatization: An overview. Manuscript in preparation.

Norcross, J. C. (1990). An eclectic definition of psychotherapy. In J. K. Zeig & W. M. Munion (Eds.), What is psychotherapy? Contemporary perspectives (pp. 218-220). San Francisco, CA: Jossey-Bass.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. Journal of Educational Statistics, 8, 157-159. doi:10.2307/1164923

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity.

Statistics in Medicine, 26(25), 4544-4562. doi:10.1002/sim.2889

R Core Team. (2019). R: A language and environment for statistical computing.

Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research? A comparative evaluation of six statistical methods. Zeitschrift für Psychologie, 227(4), 261 - 279. doi:10.31234/osf.io/w94ep

Resick, P. A., & Schnicke, M. K. (1993). Cognitive processing therapy for rape victims: A treatment manual. Newbury Park, CA: Sage.

Rhodes, K. M., Turner, R. M., & Higgins, J. P. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. Journal of Clinical Epidemiology, 68(1), 52-60.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86(3), 638-641. doi:10.1037/0033-2909.86.3.638

Rothstein, H. R., & Bushman, B. J. (2012). Publication bias in psychological science: comment on Ferguson and Brannick (2012). Psychological Methods, 17, 129-136. doi:10.1037/a0027128

Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (2nd ed., pp. 103-125). New York: Russell Sage Foundation.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis: Prevention, assessment and adjustments. Chichester, England: Wiley

Shapiro, F., & Forrest, M. S. (2001). Eye movement desensitization and reprocessing: Basic principles, protocols and procedures. New York, NY: Guilford Press.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological science, 22(11), 1359-1366. doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-Curve and effect size correcting for publication bias using only significant results. Perspectives On Psychological Science, 9(6), 666-681. doi:10.1177/1745691614553988

Sloan, D. M., Feinstein, B. A., Gallagher, M. W., Beck, J. G., & Keane, T. M. (2013). Efficacy of group treatment for posttraumatic stress disorder symptoms: A meta-analysis. Psychological Trauma: Theory, Research, Practice, and Policy, 5(2), 176-183.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. Research Synthesis Methods, 5(1), 60-78.

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. (2017). Finding the power to reduce publication bias. Statistics in Medicine. doi:10.1002/sim.7228

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. The American Statistician, 49(1), 108-112. doi:10.2307/2684823

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments (pp. 111-125). Chichester, England: Wiley.

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. Journal of clinical epidemiology, 53(11), 1119-1129.

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., . . . Schmid, C. H. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. Brithish Medical Journal, 343(7818), 1-8. doi:10.1136/bmj.d400210.3102/0162373712446144

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. Journal of clinical epidemiology, 58(9), 894-901. doi:10.1016/j.jclinepi.2005.01.006

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. Statistics in Medicine, 22(13), 2113-2126. doi:10.1002/sim.1461

Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., & Higgins, J. P. T. (2015). Predictive distributions for between-study heterogeneity and

simple methods for their application in Bayesian meta-analysis. Statistics in Medicine, 34(6), 984-998.

Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), The handbook of research synthesis and meta-analysis (Vol. 2, pp. 129-146). New York, NY: Russel Sage Foundation.

van Aert, R. C. M. (2019). puniform: Meta-Analysis Methods Correcting for Publication Bias. R package version 0.1.1. doi:https://CRAN.R-project.org/package=puniform

van Aert, R. C. M., & van Assen, M. A. L. M. (2020). Correcting for Publication Bias in a Meta-Analysis with the P-Uniform* Method. doi:10.31222/osf.io/zqjr9

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses based on p-values: Reservations and recommendations for applying p-uniform and p-curve. Perspectives On Psychological Science, 11, 713-729. doi:doi:10.1177/1745691616650874

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. Psychological Methods, 20(3), 293-309. doi:10.1037/met0000025

Veronen, L. J., & Kilpatrick, S. (1983). Stress management for rape victims. In D. Meichenbaum & M. E. Jaremko (Eds.), Stress reduction and prevention (pp. 341-374). London, England: Plenum.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. Psychometrika, 60(3), 419-435. doi:10.1007/BF02294384

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. Psychological Methods, 10(4), 428-443. doi:10.1037/1082-989X.10.4.428

Viechtbauer, W. (2007). Confidence intervals for the amount of heterogeneity in meta-analysis. Statistics in Medicine, 26(1), 37-52.

Viechtbauer, W. (2010). Conducting meta-analysis in R with the metafor package. Journal of Statistical Software, 36(3), 1-48.

Wampold, B. E., Imel, Z. E., Laska, K. M., Benish, S., Miller, S. D., Flűckiger, C., . . . Budge, S. (2010). Determining what works in the treatment of PTSD. Clinical Psychology Review, 30(8), 923-933.

Wilen, J. S. (2015). A systematic review and network meta-analysis of psychosocial interventions for adults who were sexually abused as children. (75), ProQuest Information & Learning, US. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2015-99031-072&site=ehost-live Available from EBSCOhost psyh database.

Footnotes

[1] We also coded the treatment that was studied in the meta-analyses to compare the results of the publication bias methods between the treatments. Additional information on the coding of the treatments can be found in an online repository (https://osf.io/gh729/) as well as the results split per treatment (https://osf.io/usm9f/).

[2] The four unpublished dissertations were: Bornstein, 2004; Chard, 1995; Karen, 1990; Wilen, 2015.

[3] We assumed that two-tailed hypothesis tests with α = .05 were used in the primary studies. Hence, p-uniform's effect size estimate was set equal to zero if the average of statistically significant p-value was smaller than α/4.

Appendix A

Table A

Results of traditional meta-analysis, Begg and Mazumdar's rank-correlation test, Egger's regression test, TES, p-uniform, trim and fill, PET-PEESE, and the selection model approach of Vevea and Hedges (1995) grouped by treatment category.

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (*z*)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{\mu}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
74	Jonas et al. (2013)	**TF-CBT** vs. WL / depressive symptoms (BDI, sensitivity analysis including high risk of bias studies) / post	WMD = -8.03 [-10.14, -5.93] (RE)	-8.02 [-10.07, -5.96], I^2 = 0 [0, 42.5]	7 (5)	τ = 0.24, *z* = 0.33	A = 0.01	-7.06 [-9.61, -1.33], $L^{\hat{\mu}}$ = 0.74	-8.02 [-10.07, -5.96], k^{imp} = 0	-8.68 [-11.3, -6.06]	-7.93 [-10.32, -5.53]
69	Jonas et al. (2013)	**TF-CBT** vs. WL / PTSD symptoms (sensitivity analysis including high risk of bias studies) / post	*d* = -1.13 [-1.33, -0.92] (RE)	-1.13 [-1.34, -0.92], I^2 = 0 [0, 54.2]	9 (8)	τ = 0.11, *z* = -0.27	A = 0.15	-1.11 [-1.36, -0.81], $L^{\hat{\mu}}$ = 0.19	-1.13 [-1.34, -0.92], k^{imp} = 0	-1.09 [-1.51, -0.67]	-1.11 [-1.35, -0.86]
71	Jonas et al. (2013)	**TF-CBT** vs. WL / PTSD symptoms (CAPS, sensitivity analysis including high risk of bias studies) / post	WMD = -27.21 [-32.29, -22.13] (RE)	-27.13 [-32.07, -22.2], I^2 = 0 [0, 66.4]	6 (6)	τ = 0.47, *z* = 0.93	A = 0.19	-26.3 [-31.14, -20.18], $L^{\hat{\mu}}$ = 0.33	-30.02 [-34.25, -25.8], k^{imp} = 3	-32.55 [-45.84, -19.26]	-26.78 [-34.2, -19.35]
64	Hofmann & Smits (2008)	**TF-CBT** vs. active controls / PTSD symptoms / post	*g* = 0.62 [0.28, 0.96] (RE)	0.62 [0.28, 0.97], I^2 = 48.1 [0, 92.5]	6 (3)	τ = 0.2, *z* = 0.61	A = 0	0.75 [0.15, 1.45], $L^{\hat{\mu}}$ = -0.54	0.62 [0.28, 0.97], k^{imp} = 0	-0.02 [-3.13, 3.1]	0.63 [0.13, 1.13]
2	ACPMH (2013)	**TF-CBT** vs. WL & active controls / PTSD diagnosis (ITT) / post	MH-RR = 0.51 [0.44, 0.59] (FE)	0.52 [0.46, 0.6], I^2 = 0 [0, 0]	10 (9)	τ = -0.64*, *z* = -1.9	A = 4.87*	0.56 [0.45, 0.81], $L^{\hat{\mu}}$ = 0.63	0.55 [0.48, 0.62], k^{imp} = 4	0.58 [0.48, 0.7]	0.66 [0.41, 1.08]
4	ACPMH (2013)	**TF-CBT** vs. WL & active controls / depressive symptoms (ITT) / post	*d* = -0.59 [-0.76, -0.41] (FE)	-0.59 [-0.76, -0.41], I^2 = 0 [0, 0]	11 (6)	τ = -0.02, *z* = 0.51	A = 0.19	-0.59 [-0.87, -0.1], $L^{\hat{\mu}}$ = -0.02	-0.59 [-0.76, -0.41], k^{imp} = 0	-0.64 [-1.12, -0.16]	-0.49 [-0.82, -0.16]
5	ACPMH (2013)	**TF-CBT** vs. WL & active controls / anxiety symptoms (ITT) / post	*d* = -0.64 [-0.88, -0.39] (FE)	-0.64 [-0.89, -0.39], I^2 = 0 [0, 0]	8 (4)	τ = -0.43, *z* = -2.23*	A = 0.17	-0.67 [-1.23, 0.12], $L^{\hat{\mu}}$ = -0.09	-0.47 [-0.69, -0.25], k^{imp} = 3	0.84 [-0.26, 1.93]	-0.54 [-1.2, 0.13]
6	ACPMH (2013)	TF-CBT vs. **WL & active controls** / attrition (ITT) / post	MH-RR = 1.48 [0.99, 2.21] (FE)	1.29 [0.84, 1.98], I^2 = 0 [0, 0]	10 (1)	τ = 0.02, *z* = 0.71	A = 0.42	0^0 [0, 0.8], $L^{\hat{\mu}}$ = 1.66*	1.29 [0.84, 1.98], k^{imp} = 0	1 [0.34, 2.92]	1.25 [0.64, 2.42]
14	Bisson et al. (2007) b	TF-CBT vs. **WL & active controls** / withdrawal rate / post	MH-RR = 1.42 [1.05, 1.94] (FE)	1.3 [0.94, 1.78], I^2 = 0 [0, 0]	15 (0)	τ = -0.24, *z* = 1.12	A = 0.93	No significant studies	1.16 [0.86, 1.57], k^{imp} = 4	0.97 [0.52, 1.84]	1.38 [0.98, 1.94]

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (z)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{\mu}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
17	Bisson et al. (2013)	TF-CBT vs. **WL & active controls** / leaving the study early / post	MH-RR = 1.64 [1.30, 2.06] (FE)	1.45 [1.15, 1.84], I^2 = 0 [0, 0]	29 (2)	τ = −0.09, *z* = 1.24	A = 0.06	0.05^0 [0, 9.1], $L^{\hat{\mu}}$ = 1.07	1.32 [1.05, 1.65], k^{imp} = 4	1.12 [0.68, 1.83]	1.92 [1.12, 3.29]
29	Bisson et al. (2013)	TF-CBT (individual & group) vs. **WL & active controls** / leaving the study early / post	MH-RR = 1.21 [0.94, 1.55] (FE)	1.19 [0.93, 1.52], I^2 = 0 [0, 0]	7 (0)	τ = 0.24, *z* = 0.44	A = 0.54	No significant studies	1.19 [0.93, 1.52], k^{imp} = 0	1.1 [0.76, 1.59]	1.18 [NaN, NaN]
97	Bisson et al. (2013)	**TF-CBT** vs. WL & active controls / anxiety / post	SMD = −0.81 [−1.03, −0.59] (RE)	−0.8 [−1.02, −0.58], I^2=43.3 [0, 71.3]	17 (10)	τ = −0.06, *z* = −0.76	A = 0.04	−0.95 [−1.22, −0.61], $L^{\hat{\mu}}$ = −1.09	−0.8 [−1.02, −0.58], k^{imp} = 0	−0.18 [−1.38, 1.02]	−0.82 [−1.15, −0.49]
67	Jonas et al. (2013)	**TF-CBT** vs. WL & active controls / PTSD symptoms / post	*d* = −1.27 [−1.54, −1.00] (RE)	−1.27 [−1.54, −1], I^2 = 23.7 [0, 85.7]	7 (7)	τ = −0.33, *z* = −1.84	A = 0.73	−1.29 [−1.6, −1.01], $L^{\hat{\mu}}$ = −0.39	−1.11 [−1.4, −0.82], k^{imp} = 3	−0.53 [−1.51, 0.45]	−1.2 [−1.49, −0.92]
68	Jonas et al. (2013)	**TF-CBT** vs. WL & active controls / PTSD symptoms (sensitivity analysis including high risk of bias studies) / post	*d* = −1.19 [−1.38, −0.99] (RE)	−1.2 [−1.4, −1], I^2 = 0 [0, 68.5]	11 (10)	τ = −0.09, *z* = −1.09	A = 0.31	−1.2 [−1.46 ,−0.93], $L^{\hat{\mu}}$ = −0.04	−1.12 [−1.34, −0.89], k^{imp} = 2	−1.01 [−1.45, −0.57]	−1.16 [−1.39, −0.93]
70	Jonas et al. (2013)	**TF-CBT** vs. WL & active controls / PTSD symptoms (CAPS, sensitivity analysis including high risk of bias studies) / post	WMD = −27.92 [−32.87, −22.96] (RE)	−27.88 [−32.68, −23.07], I^2 = 0 [0, 69.5]	7 (7)	τ = 0.05, *z* = −0.13	A = 0.43	−27.6 [−32.55, −21.87], $L^{\hat{\mu}}$ = 0.11	−27.88 [−32.68, −23.07], k^{imp} = 0	−26.1 [−38.42, −13.78]	No convergence
72	Jonas et al. (2013)	**TF-CBT** vs. WL & active controls / depressive symptoms (BDI) / post	WMD = −8.21 [−10.30, −6.12] (RE)	−8.21 [−10.25, −6.17], I^2 = 0 [0, 29.7]	6 (6)	τ = −0.2, *z* = −0.27	A = 1.36	−6.93 [−9.32, −2.52], $L^{\hat{\mu}}$ = 1.05	−8.21 [−10.25, −6.17], k^{imp} = 0	−7.79 [−11.21, −4.38]	−7.3 [−9.26, −5.33]
73	Jonas et al. (2013)	**TF-CBT** vs. WL & active controls / depressive symptoms (BDI, sensitivity analysis including high risk of bias studies) / post	WMD = −7.85 [−9.80, −5.89] (RE)	−7.82 [−9.72, −5.92], I^2 = 0 [0, 32.8]	9 (6)	τ = 0.39, *z* = 0.69	A = 0	−6.93 [−9.32, −2.52], $L^{\hat{\mu}}$ = 0.72	−8.04 [−9.89, −6.18], k^{imp} = 1	−8.97 [−11.21, −6.73]	−7.75 [−10, −5.51]
75	Jonas et al. (2013)	**TF-CBT** vs. WL & active controls (including PCT) / depressive symptoms (BDI, sensitivity analysis including PCT) / post	WMD = −6.91 [−8.86, −4.96] (RE)	−6.96 [−8.91, −5.01], I^2 = 23.2 [0, 75.6]	7 (7)	τ = −0.33, *z* = −2.08*	A = 2.84	−6.02 [−8.63, −2.76], $L^{\hat{\mu}}$ = 0.38	−5.86 [−7.79, −3.93], k^{imp} = 3	−1.97 [−7.02, 3.07]	−5.2 [−9.42, −0.98]

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (z)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{\mu}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
76	Jonas et al. (2013)	**TF-CBT** vs. WL & active controls (including PCT) / depressive symptoms (BDI, sensitivity analysis including PCT and high risk of bias studies) / post	WMD = -6.29 [-7.84, -4.75] (RE)	-6.38 [-7.99, -4.76], I^2 = 6.4 [0, 68.7]	11 (7)	τ = 0.31, *z* = -0.12	A = 0.45	-6.02 [-8.63, -2.76], $L^{\hat{\mu}}$ = 0.17	-6.38 [-7.99, -4.76], k^{imp} = 0	-6.26 [-9.01, -3.52]	-5.92 [-8.17, -3.67]
45	DiMauro (2014)	TF-CBT within group / PTSD symptoms / post	*d* = 0.69 [0.35, 1.02] (RE)	0.68 [0.27, 1.09], I^2 = 0 [0, 95.2]	6 (1)	τ = 0.6, *z* = 1.6	A = 0.26	-6.58^0 [-51.19, 5.44], $L^{\hat{\mu}}$ = 1.01	0.6 [0.21, 0.99], k^{imp} = 2	-0.1 [-0.78, 0.58]	0.87 [NaN, NaN]
8	ACPMH (2013)	**CBT combined** (mostly TF-CBT, individual & group) vs. active controls / PTSD symptoms (clinician rated) / 2-3 months follow-up	*d* = -0.43 [-0.65, -0.20] (FE)	-0.43 [-0.65, -0.2], I^2 = 0 [0, 0]	7 (2)	τ = -0.33, *z* = -1.58	A = 0	-0.34 [-1.3, 1.26], $L^{\hat{\mu}}$ = 0.19	-0.38 [-0.6, -0.17], k^{imp} = 1	0.42 [-0.68, 1.53]	-0.44 [-0.61, -0.26]
9	ACPMH (2013)	**CBT combined (mostly TF-CBT)** vs. active controls / depressive symptoms / post	*d* = -0.68 [-0.92, -0.44] (FE)	-0.68 [-0.92, -0.45], I^2 = 0 [0, 0]	8 (4)	τ = -0.29, *z* = -1.33	A = 0.01	-0.59 [-1.18, 1.14], $L^{\hat{\mu}}$ = 0.32	-0.56 [-0.77, -0.35], k^{imp} = 3	-0.12 [-1.01, 0.77]	-0.67 [-1.14, -0.19]
10	ACPMH (2013)	CBT combined (mostly TF-CBT, individual & group) vs. **active controls** / attrition / post	MH-RR = 1.36 [0.86, 2.15] (FE)	1.27 [0.79, 2.05], I^2 = 0 [0, 0]	10 (0)	τ = 0.38, *z* = 1.15	A = 0.51	No significant studies	1.18 [0.74, 1.87], k^{imp} = 2	0.76 [0.34, 1.7]	1.25 [NaN, NaN]
1	ACPMH (2013)	**CBT combined** (mostly TF-CBT) vs. WL & active controls / PTSD symptoms (self-rated) / post	*d* = -1.14 [-1.32, -0.95] (FE)	-1.14 [-1.32, -0.95], I^2 = 0 [0, 0]	11 (10)	τ = -0.24, *z* = -1.11	A = 0.47	-1.12 [-1.37, -0.9], $L^{\hat{\mu}}$ = 0.14	-1.14 [-1.32, -0.95], k^{imp} = 0	-1.03 [-1.39, -0.68]	-1.11 [-1.31, -0.9]
3	ACPMH (2013)	**CBT combined** (mostly TF-CBT) vs. WL & active controls / PTSD symptoms (self-rated, ITT) / post	*d* = -1.06 [-1.30, -0.82] (FE)	-1.08 [-1.32, -0.84], I^2 = 0 [0, 0]	6 (5)	τ = -0.07, *z* = -0.7	A = 0.15	-1.03 [-1.37, -0.71], $L^{\hat{\mu}}$ = 0.37	-1.08 [-1.32, -0.84], k^{imp} = 0	-1.01 [-1.6, -0.43]	-1.05 [-1.31, -0.8]
7	ACPMH (2013)	**CBT combined** (mostly TF-CBT, individual & group) vs. WL & active controls / PTSD symptoms (self-rated, motor vehicle accident) / post	*d* = -1.25 [-1.57, -0.94] (FE)	-1.26 [-1.57, -0.94], I^2 = 0 [0, 0]	6 (5)	τ = 0.33, *z* = 0.6	A = 0.03	-1.22 [-1.59, -0.65], $L^{\hat{\mu}}$ = 0.18	-1.29 [-1.59, -0.98], k^{imp} = 1	-1.4 [-2.13, -0.68]	-1.24 [-1.5, -0.97]
30	Casement & Swanson (2012)	CBT combined (mostly TF-CBT) within group / nightmare frequency / post	*g* = 0.82 [0.57, 1.07] (RE)	0.82 [0.66, 0.98], I^2 = 2.8 [0, 83.3]	7 (6)	τ = -0.33, *z* = -0.13	A = 0	0.75 [0.45, 0.95], $L^{\hat{\mu}}$ = 0.77	0.86 [0.72, 1.01], k^{imp} = 2	0.84 [0.46, 1.22]	0.82 [0.66, 0.99]

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (z)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{\mu}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
31	Casement & Swanson (2012)	CBT combined (mostly TF-CBT) within group / PTSD symptoms / post	g = 0.71 [0.46, 0.95] (RE)	0.69 [0.5, 0.88], I^2 = 28.4 [0, 91]	7 (5)	τ = 0.24, z = 1.12	A = 0.05	0.77 [0.55, 1.05], $L^{\hat{\mu}}$ = −0.91	0.66 [0.46, 0.86], k^{imp} = 1	0.56 [0.23, 0.88]	0.68 [0.49, 0.88]
48	Dorrepaal et al. (2014)	CBT combined (mostly TF-CBT, individual & group) within group / PTSD symptoms (completer) / post	d = 1.7 (FE)	1.68 [1.36, 2], I^2 = 0 [0, 0]	8 (8)	τ = 0.43, z = 2.76*	A = 0.78	1.86 [1.34, 2.27], $L^{\hat{\mu}}$ = −0.74	1.43 [1.14, 1.71], k^{imp} = 3	−0.55 [−2.23, 1.14]	1.56 [1.24, 1.87]
49	Dorrepaal et al. (2014)	CBT combined (mostly TF-CBT, individual & group) within group / PTSD symptoms (ITT) / post	d = 1.3 (FE)	1.29 [1.05, 1.52], I^2 = 0 [0, 0]	8 (7)	τ = 0.64*, z = 2.73*	A = 0.98	1.45 [1.14, 1.75], $L^{\hat{\mu}}$ = −1.07	1.15 [0.94, 1.37], k^{imp} = 2	−1.66 [−3.8, 0.49]	1.36 [1.06, 1.66]
52	Dorrepaal et al. (2014)	CBT combined (mostly TF-CBT, individual & group) within group / PTSD symptoms (completer, complex PTSD) / post	d = 1.6 (FE)	1.54 [1.18, 1.9], I^2 = 0 [0, 0]	6 (6)	τ = 0.47, z = 2.31*	A = 0.82	1.64 [1.03, 2.18], $L^{\hat{\mu}}$ = −0.34	1.17 [0.86, 1.47], k^{imp} = 3	−0.46 [−2.71, 1.78]	1.33 [0.98, 1.68]
53	Dorrepaal et al. (2014)	CBT combined (mostly TF-CBT, individual & group) within group / PTSD symptoms (ITT, complex PTSD) / post	d = 1.2 (FE)	1.17 [0.91, 1.44], I^2 = 0 [0, 0]	6 (5)	τ = 0.73, z = 1.99*	A = 0.96	1.31 [0.97, 1.67], $L^{\hat{\mu}}$ = −0.8	1.08 [0.84, 1.33], k^{imp} = 1	−1.51 [−5.15, 2.13]	1.26 [0.95, 1.57]
32	Chen et al. (2014)	**EMDR** vs. active controls (including PE, SIT/PE) / depressive symptoms / post	g = −0.45 [−0.65, −0.25] (RE)	−0.45 [−0.65, −0.25], I^2 = 0 [0, 63.3]	11 (3)	τ = −0.2, z = −0.83	A = 0	−0.22□ [−1.25, 2.67], $L^{\hat{\mu}}$ = 0.39	−0.45□ [−0.65, −0.25], k^{imp} = 0	−0.07□ [−0.96, 0.82]	−0.44 [−0.71, −0.18]
34	Chen et al. (2014)	**EMDR** vs. active controls (including TTP) / anxiety symptoms (equivalent group) / post	g = −0.41 [−0.62, −0.21] (RE)	−0.41 [−0.62, −0.2], I^2 = 0 [0, 74.5]	8 (4)	τ = −0.21, z = 0	A = 2.11	$0.34^{0,□}$ [−0.39, 1.93], $L^{\hat{\mu}}$ = 2*	−0.41□ [−0.62, −0.2], k^{imp} = 0	−0.41□ [−1.16, 0.34]	−0.16 [NaN, NaN]
35	Chen et al. (2014)	**EMDR** vs. active controls (including TTP, Exp) / subjective distress (equivalent group) / post	g = −0.57 [−0.81, −0.33] (RE)	−0.57 [−0.81, −0.33], I^2 = 0 [0, 64]	8 (2)	τ = −0.14, z = 0.53	A = 0.47	−0.43□ [−1.19, 26.29], $L^{\hat{\mu}}$ = 0.13	−0.57□ [−0.81, −0.33], k^{imp} = 0	−0.68□ [−1.3, −0.05]	−0.73 [−1.09, −0.36]
37	Chen et al. (2014)	**EMDR** vs. active controls (including TTP, PE, CBT, SIT/PE, Exp, SM) / PTSD symptoms (equivalent group) / post	g = −0.58 [−0.73, −0.42] (RE)	−0.57 [−0.73, −0.42], I^2 = 3 [0, 63.3]	13 (5)	τ = 0.1, z = 0.56	A = 0.64	−0.67□ [−0.94, −0.34], $L^{\hat{\mu}}$ = −0.76	−0.7□ [−0.87, −0.52], k^{imp} = 4	−0.63□ [−0.94, −0.32]	−0.66 [−0.87, −0.46]

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (z)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{R}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
15	Bisson et al. (2007) b	EMDR vs. **WL & active controls** / withdrawal rate / post	MH-RR = 1.21 [0.66, 2.22] (RE)	1.27 [0.69, 2.35], I^2 = 0 [0, 55.1]	6 (0)	τ = −0.07, *z* = −0.88	A = 0.31	No significant studies	1.66 [1, 2.77], k^{imp} = 3	1.66 [0.8, 3.43]	1.25 [0.77, 2]
24	Bisson et al. (2013)	EMDR vs. **WL & active controls** / leaving the study early / post	MH-RR = 1.05 [0.62, 1.79] (FE)	1.04 [0.6, 1.8], I^2 = 0 [0, 0]	7 (0)	τ = 0.14, *z* = −0.77	A = 0.2	No significant studies	1.19 [0.71, 2], k^{imp} = 2	1.64 [0.51, 5.25]	1 [0.68, 1.47]
25	Bisson et al. (2013)	**EMDR** vs. WL & active controls / depressive symptoms / post	*d* = -1.15 [-1.52, -0.78] (RE)	-1.15 [-1.52, -0.78], I^2 = 38.3 [0, 88.2]	7 (5)	τ = 0.2, *z* = 0.15	A = 0.5	-1.32 [-1.71, -0.91], $L^{\hat{R}}$ = −0.82	-1.15 [-1.52, -0.78], k^{imp} = 0	-1.47 [-4.8, 1.86]	-1.26 [-1.63, -0.88]
26	Bisson et al. (2013)	**EMDR** vs. WL & active controls / anxiety symptoms / post	*d* = -1.02 [-1.36, -0.69] (FE)	-1.02 [-1.35, -0.69], I^2 = 0 [0, 0]	6 (5)	τ = 0.2, *z* = 0.94	A = 0.78	-0.19 [-1.14, 1.95], $L^{\hat{R}}$ = 1.6	-1.02 [-1.35, -0.69], k^{imp} = 0	-1.49 [-2.71, -0.27]	-0.62 [-1.64, 0.39]
33	Chen et al. (2014)	**EMDR** vs. WL & active controls / depressive symptoms (adults) / post	*g* = -0.63 [-0.83, -0.44] (RE)	-0.64 [-0.83, -0.44], I^2 = 37.1 [0, 69.4]	17 (6)	τ = −0.09, *z* = 0.21	A = 1.51	-0.78□ [-1.27, -0.11], $L^{\hat{R}}$ = −0.45	-0.64□ [-0.83, -0.44], k^{imp} = 0	-0.72□ [-1.17, -0.26]	-0.83 [-1.12, -0.54]
36	Chen et al. (2014)	**EMDR** vs. WL & active controls (including CBT, Exposure) / PTSD symptoms (<60 min/session) / post	*g* = -0.50 [-0.74, -0.27] (RE)	-0.5 [-0.74, -0.26], I^2 = 35.6 [0, 78.7]	10[c] (4)	τ = −0.2, *z* = −0.18	A = 0.01	-0.57□ [-1.02, 0.64], $L^{\hat{R}}$ = −0.27	-0.5□ [-0.74, -0.26], k^{imp} = 0	-0.5□ [-1.4, 0.4]	-0.48 [-0.84, -0.13]
38	Chen et al. (2014)	**EMDR** vs. WL & active controls (including TTP, SIT/PE, SM) / depressive symptoms (with manual) / post	*g* = -0.55 [-0.74, -0.36] (RE)	-0.55 [-0.74, -0.36], I^2 = 35.7 [0, 66.9]	18[c] (5)	τ = −0.08, *z* = 0.53	A = 1.39	-0.62□ [-1.26, 0.27], $L^{\hat{R}}$ = −0.13	-0.6□ [-0.79, -0.41], k^{imp} = 2	-0.7□ [-1.14, -0.26]	-0.75 [-1.05, -0.44]
39	Chen et al. (2014)	**EMDR** vs. WL & active controls / depressive symptoms (<60 min/session) / post	*g* = -0.30 [-0.55, -0.04] (RE)	-0.3 [-0.55, -0.04], I^2 = 0 [0, 44.8]	6[c] (0)	τ = 0.47, *z* = 0.74	A = 1.04	No significant studies	-0.4□ [-0.62, -0.18], k^{imp} = 3	-0.59□ [-1.13, -0.04]	-0.45 [-0.63, -0.27]
40	Chen et al. (2014)	**EMDR** vs. WL & active controls / anxiety symptoms (<60 min/session) / post	*g* = -0.35 [-0.58, -0.13] (RE)	-0.35 [-0.57, -0.13], I^2 = 0 [0, 88.4]	6[c] (3)	τ = −0.07, *z* = −0.18	A = 2.34	$0.5^{0,□}$ [-0.45, 3.24], $L^{\hat{R}}$ = 1.72*	-0.35□ [-0.57, -0.13], k^{imp} = 0	-0.31□ [-1.41, 0.79]	-0.08 [-0.63, 0.47]
61	Gerger et al. (2014) b	Other therapies within group / PTSD symptoms (non-complex problems) / post	*g* = -0.71 [-1.02, -0.40] (RE)	-0.71 [-1.02, -0.4], I^2 = 42.5 [0, 88.4]	6 (4)	τ = −0.6, *z* = -1.97*	A = 0.08	-0.65□ [-1.31, -0.13], $L^{\hat{R}}$ = 0.08	-0.71□ [-1.02, -0.4], k^{imp} = 0	0.35□ [-1.16, 1.87]	-0.62 [-1.03, -0.2]
81	Peleikis & Dahl (2005)	**Combined therapies** (mostly other therapies, group) vs. WL / trauma symptoms / post	*d* = 0.44 [0.25, 0.64] (*w*=*n*)	0.43 [0.23, 0.63], I^2 = 0 [0, 0]	8 (2)	τ = 0.14, *z* = 1.54	A = 0.13	0.91 [-0.58, 1.57], $L^{\hat{R}}$ = −0.98	0.43 [0.23, 0.63], k^{imp} = 0	-0.18 [-1.41, 1.05]	0.66 [0.08, 1.24]

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (z)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{\mu}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
87	Sloan et al. (2013)	**Combined therapies** (mostly CBT, group) vs. WL / PTSD symptoms / post	g = 0.56 [0.31, 0.82] (RE)	0.56 [0.32, 0.79], I^2 = 0 [0, 71]	6 (3)	τ = 0.6, z = 1.27	A = 0.02	0.21 [-1.27, 0.9], $L^{\hat{\mu}}$ = 0.93	0.5 [0.28, 0.72], k^{imp} = 1	-0.95 [-2.85, 0.95	0.53 [0.26, 0.79]
89	Taylor & Harvey (2009)	**Combined therapies** (mostly TF-CBT) vs. WL / mixed outcome measures (7-9 sessions) / post	g = 0.89 [0.58, 1.21] (FE)	0.89 [0.58, 1.21], I^2 = 0 [0, 0]	6 (4)	τ = -0.2, z = -1.29	A = 0.04	0.88 [-0.28, 1.46], $L^{\hat{\mu}}$ = 0.03	0.89 [0.58, 1.21], k^{imp} = 0	3.12 [-3.23, 9.48]	0.72 [-0.01, 1.45]
90	Taylor & Harvey (2009)	**Combined therapies** (mostly TF-CBT) vs. WL / mixed outcome measures (practitioner as therapist) / post	g = 0.98 [0.70, 1.26] (FE)	0.99 [0.71, 1.26], I^2 = 0 [0, 0]	7 (5)	τ = 0.14, z = -0.34	A = 0.02	1.04 [-0.18, 1.43], $L^{\hat{\mu}}$ = -0.24	0.99 [0.71, 1.26], k^{imp} = 0	1 [-0.11, 2.11]	0.86 [0.2, 1.51]
78	Lambert & Alhassoon (2014) b	**Combined therapies** (mostly TF-CBT, individual & group) vs. active controls (including SIT) / depressive symptoms / post	g = 0.63 [0.35, 0.92] (RE)	0.69 [0.36, 1.03], I^2 = 28.7 [0, 86.3]	9 (2)	τ = 0.56*, z = 2.23*	A = 0.46	1.77 [0.6, 2.6], $L^{\hat{\mu}}$ = -1.87	0.69 [0.36, 1.03], k^{imp} = 0	-0.36 [-1.42, 0.69]	1.19 [0.2, 2.18]
88	Sloan et al. (2013)	**Combined therapies** (mostly CBT, group) vs. active controls / PTSD symptoms / post	g = 0.09 [-0.03, 0.22] (RE)	0.15 [0, 0.3], I^2 = 39.4 [0, 94.6]	10 (2)	τ = 0.47, z = 2.66*	A = 1.18	0.38 [-3.25, 1.59], $L^{\hat{\mu}}$ = -0.3	0.1 [-0.08, 0.28], k^{imp} = 2	-0.18 [-0.47, 0.1]	0.06 [-0.06, 0.18]
79	Nenova et al. (2013)	**Combined therapies** (including SN, individual & group, combined delivery) vs. WL & active controls / PTSD symptoms (intrusion) / post	delta = -0.09 [-0.41, 0.26] (Bayesian RE)	-0.1 [-0.32, 0.11], I^2 = 0 [0, 0]	13 (0)	τ = -0.18, z = 0.11	A = 0.56	No significant studies	-0.1 [-0.32, 0.11], k^{imp} = 0	-0.1 [-0.21, 0.01]	-0.1 [-0.26, 0.07]
80	Nenova et al. (2013)	Combined therapies (including SN, individual & group, combined delivery) vs. WL & active controls / PTSD symptoms (avoidance) / post	delta = 0.00 [-0.37, 0.31] (Bayesian RE)	0.04 [-0.13, 0.22], I^2 = 0 [0, 0]	13 (0)	τ = -0.13, z = -0.66	A = 0.41	No significant studies	0.06 [-0.12, 0.23], k^{imp} = 6	0.06 [0.02, 0.1]	No convergence
84	Sherman (1998)	**Combined therapies** (individual & group) vs. WL & active controls (one study removed) / mixed outcomes / post	g = 0.52 [0.37, 0.67][b]	0.52 [0.38, 0.66], I^2 = 0 [0, 0]	23 (4)	τ = 0.37*, z = 1.54	A = 2.06	0.62 [-0.35, 1.2], $L^{\hat{\mu}}$ = -0.28	0.42 [0.29, 0.55], k^{imp} = 6	-0.11 [-0.71, 0.5]	0.72 [0.1, 1.35]
85	Sherman (1998)	**Combined therapies** (individual & group) vs. WL & active controls / mixed	g = 0.64 [0.47, 0.81][b]	0.64 [0.48, 0.81], I^2 = 0 [0, 0]	19 (10)	τ = 0.3, z = 1.74	A = 0.97	0.7 [0.25, 1.03], $L^{\hat{\mu}}$ = -0.28	0.59 [0.43, 0.75], k^{imp} = 2	0.02 [-0.84, 0.89]	0.49 [0.22, 0.76]

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (z)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{\mu}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
		outcome measures (one study removed) / follow-up									
86	Sloan et al. (2013)	**Combined therapies** (mostly CBT, group) vs. WL & active controls / PTSD symptoms / post	*g* = 0.24 [0.09, 0.39] (RE)	0.28 [0.13, 0.43], I^2 = 48.8 [8, 87]	16 (5)	τ = 0.65*, *z* = 4.33*	A = 2.92	0.26 [-0.9, 0.89], $L^{\hat{\mu}}$ = -0.17	0.13 [-0.04, 0.3], k^{imp} = 6	-0.25 [-0.46, -0.04]	0.09 [-0.01, 0.19]
93	Tol et al. (2011)	**Combined therapies** vs. WL & active controls / PTSD symptoms / post	*g* = -0.38 [-0.55, -0.20] (RE)	-0.38 [-0.56, -0.2], I^2 = 22.1 [0, 78.4]	9 (3)	τ = -0.39, *z* = -1.77	A = 0.01	-0.56 [-0.9, -0.06], $L^{\hat{\mu}}$ = -0.99	-0.35 [-0.53, -0.17], k^{imp} = 1	0.34 [-0.62, 1.29]	-0.43 [-0.75, -0.11]
46	Dorrepaal et al. (2014)	Combined therapies (mostly CBT, individual & group) within group / PTSD symptoms (completer) / post	*d* = 1.7 (FE)	1.65 [1.35, 1.95], I^2 = 0 [0, 0]	9 (9)	τ = 0.5, *z* = 2.78*	A = 0.92	1.77 [1.3, 2.17], $L^{\hat{\mu}}$ = -0.54	1.34 [1.08, 1.6], k^{imp} = 4	-0.6 [-2.14, 0.95]	1.51 [1.23, 1.79]
47	Dorrepaal et al. (2014)	Combined therapies (mostly TF-CBT, individual & group) within group / PTSD symptoms (ITT) / post	*d* = 1.3 (FE)	1.28 [1.06, 1.51], I^2 = 0 [, 0]	9 (8)	τ = 0.44, *z* = 2.44*	A = 0.5	1.41 [1.12, 1.69], $L^{\hat{\mu}}$ = -0.91	1.17 [0.96, 1.37], k^{imp} = 2	-1.12 [-3.19, 0.96]	1.34 [1.06, 1.62]
50	Dorrepaal et al. (2014)	Combined therapies (mostly CBT, individual & group) within group / PTSD symptoms (completer, complex PTSD) / post	*d* = 1.6 (FE)	1.52 [1.19, 1.85], I^2 = 0 [0, 0]	7 (7)	τ = 0.62, *z* = 2.3*	A = 0.96	1.56 [1.05, 2.06], $L^{\hat{\mu}}$ = -0.17	1.34 [1.04, 1.64], k^{imp} = 2	-0.48 [-2.38, 1.41]	1.33 [1.03, 1.63]
51	Dorrepaal et al. (2014)	Combined therapies (mostly CBT) within group / PTSD symptoms (ITT, complex PTSD) / post	*d* = 1.2 (FE)	1.18 [0.93, 1.43], I^2 = 0 [0, 0]	7 (6)	τ = 0.52, *z* = 1.72	A = 0.4	1.28 [0.97, 1.6], $L^{\hat{\mu}}$ = -0.67	1.18 [0.93, 1.43], k^{imp} = 0	-0.74 [-3.55, 2.08]	1.24 [0.95, 1.53]
106	Ehring et al. (2014)	Combined therapies (trauma-focused) within group / PTSD symptoms / follow-up	*g* = 1.83 [1.60, 2.09] (RE)	1.85 [1.53, 2.17], I^2 = 22.6 [0, 84]	7 (7)	τ = 0.43, *z* = 2.12*	A = 0.1	1.89 [1.56, 2.26], $L^{\hat{\mu}}$ = -0.4	1.79 [1.46, 2.11] , k^{imp} = 1	0.36 [-1.1, 1.83]	1.83 [1.55, 2.11]
91	Taylor & Harvey (2009)	Combined therapies (mostly TF-CBT) within group / mixed outcome measures / post	*g* = 1.11 [0.90, 1.32] (FE)	1.08 [0.91, 1.25], I^2 = 0 [0, 0]	6 (6)	τ = 0.6, *z* = 2.49*	A = 1.34	1.16 [0.97, 1.66], $L^{\hat{\mu}}$ = -0.75	1.02 [0.86, 1.19], k^{imp} = 2	0.91 [0.77, 1.05]	1.05 [0.89, 1.21]
92	Taylor & Harvey (2009)	Combined therapies (mostly TF-CBT, individual & group) within group / mixed outcome measures (therapist as main contact for assessment and treatment) / post	*g* = 1.03 [0.83, 1.23] (FE)	1 [0.84, 1.17], I^2 = 0 [0, 0]	7 (6)	τ = 0.14, *z* = 1.13	A = 0.06	1.06 [0.88, 1.35], $L^{\hat{\mu}}$ = -0.65	1 [0.84, 1.17], k^{imp} = 0	0.89 [0.58, 1.2]	0.99 [0.82, 1.17]

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (z)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{\mu}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
43	Diehle et al. (2014)	**TF-CBT (cognitive restructuring + exposure)** vs. TF-CBT (exposure only) / trauma-related cognitions / post	*g* = 0.27 [0.03, 0.50] (RE)	0.26 [0.02, 0.5], I^2 = 18.2 [0, 85.8]	7 (1)	τ = −0.43, *z* = −1.37	A = 0	0.62 [−0.32, 1.18], $L^{\hat{\mu}}$ = −1.08	0.3 [0.05, 0.56], k^{imp} = 1	0.59 [−0.06, 1.25]	0.28 [−0.02, 0.58]
44	Diehle et al. (2014)	**TF-CBT (cognitive restructuring + exposure)** vs. TF-CBT (exposure only) / trauma-related cognitions / follow-up	*g* = 0.15 [−0.08, 0.39] (RE)	0.15 [−0.08, 0.39], I^2 = 16.1 [0, 87]	7 (0)	τ = −0.14, *z* = −0.57	A = 0.58	No significant studies	0.19 [−0.05, 0.43], k^{imp} = 1	0.44 [−1.08, 1.96]	No convergence
77	Kehle-Forbes et al. (2013)	**TF-CBT (exposure only)** vs. TF-CBT (exposure plus) / dropout / post	RR = 0.97 [0.66, 1.41] (RE)	0.97 [0.66, 1.41], I^2 = 40.5 [0, 92.2]	8 (1)	τ = −0.36, *z* = −0.82	A = 0.3	0^0 [0, 1.91], $L^{\hat{\mu}}$ = 1.5	1.09 [0.73, 1.62], k^{imp} = 2	1 [0.41, 2.4]	0.87 [0.71, 1.07]
18	Bisson et al. (2013)	TF-CBT vs. **Non-TF-CBT** / leaving the study early / post	MH-RR = 1.19 [0.71, 2.00] (FE)	1.12 [0.67, 1.89], I^2 = 0 [0, 0]	7 (0)	τ = −0.05, *z* = 0.67	A = 0.26	No significant studies	1.05 [0.63, 1.74], k^{imp} = 1	0.76 [0.24, 2.41]	1.2 [0.5, 2.88]
19	Bisson et al. (2013)	**TF-CBT** vs. Non-TF-CBT / depressive symptoms / post	*d* = −0.27 [−0.56, 0.03] (FE)	-0.27 [−0.56, 0.03], I^2 = 0 [0, 0]	6 (0)	τ = 0.2, *z* = 0.08	A = 0.72	No significant studies	-0.27 [−0.56, 0.03], k^{imp} = 0	-0.33 [−2.85, 2.19]	-0.26 [−0.56, 0.03]
20	Bisson et al. (2013)	**TF-CBT** vs. Non-TF-CBT / PTSD diagnosis / post	MH-RR = 0.83 [0.60, 1.17] (RE)	0.84 [0.61, 1.16], I^2 = 36.9 [0, 97.6]	6 (1)	τ = −0.47, *z* = −1.82	A = 0.83	0.6 [0.02, 3549.05], $L^{\hat{\mu}}$ = -0.18	0.84 [0.61, 1.16], k^{imp} = 0	1.56 [0.55, 4.38]	0.93 [0.72, 1.22]
98	Bisson et al. (2013)	TF-CBT vs. Non-TF-CBT / PTSD diagnosis clinician rated / post	SMD = -0.27 [-0.63, 0.10] (RE)	-0.26 [−0.62, 0.1], I^2 = 48.1 [0, 92.8]	7 (1)	τ = 0.05, Z = −0.82	A = 0.1	-1.19 [−2.3, 0.68], $L^{\hat{\mu}}$ = -1.3	-0.26 [−0.62, 0.1], k^{imp} = 0	0.32 [−1.99, 2.63]	-0.3 [−1.36, 0.75]
21	Bisson et al. (2013)	TF-CBT vs. **Other therapies (including TAU)** / leaving the study early / post	MH-RR = 1.39 [1.01, 1.92] (FE)	1.36 [0.98, 1.87], I^2 = 0 [0, 97.6]	11 (1)	τ = −0.06, *z* = −0.03	A = 0.03	0.31^0 [0, 12.22], $L^{\hat{\mu}}$ = 0.69	1.38 [1.01, 1.9], k^{imp} = 2	1.36 [0.76, 2.46]	1.33 [0.93, 1.9]
22	Bisson et al. (2013)	**TF-CBT** vs. Other therapies (including TAU) / depressive symptoms (self-rated) / post	*d* = −0.37 [−0.63, -0.11] (RE)	-0.38 [−0.64, -0.11], I^2 = 42.7 [0, 90.4]	9 (3)	τ = −0.39, *z* = −1.73	A = 1.11	-0.33 [-1.04, 2.1], $L^{\hat{\mu}}$ = -0.07	-0.26 [−0.56, 0.03], k^{imp} = 2	0.01 [−0.48, 0.49]	-0.18 [−0.39, 0.02]
23	Bisson et al. (2013)	**TF-CBT** vs. Other therapies (including TAU) / PTSD diagnosis / post	MH-RR = 0.75 [0.59, 0.96] (RE)	0.76 [0.6, 0.96], I^2 = 34.3 [0, 91.5]	7 (1)	τ = −0.52, *z* = −2.31*	A = 0.03	0.71 [0.17, 25.4], $L^{\hat{\mu}}$ = -0.12	0.83 [0.64, 1.08], k^{imp} = 2	1.53 [0.77, 3.03]	0.52 [0.18, 1.48]
57	Gerger et al. (2014) b	**TF-CBT** vs. Other therapies / PTSD symptoms (structural equivalence) / post	*g* = −0.17 [−0.39, 0.06] (RE)	-0.16 [−0.38, 0.06], I^2 = 44.3 [0, 89.2]	7 (2)	τ = −0.05, *z* = 0.18	A = 1.03	-0.34 [-0.71, 0.36], $L^{\hat{\mu}}$ = -1	-0.16 [−0.38, 0.06], k^{imp} = 0	-0.26 [−0.94, 0.41]	-0.06 [-0.3, 0.18]

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (z)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{\mu}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
58	Gerger et al. (2014) b	**TF-CBT** vs. Other therapies / PTSD symptoms (complex problem & structural equivalence) / post	*g* = -0.11 [-0.32, 0.09] (RE)	-0.11 [-0.31, 0.09], I^2 = 33.5 [0, 83.6]	6 (1)	τ = -0.07, *z* = 0.88	A = 0.14	-0.36 [-0.61, 0.02], $L^{\hat{\mu}}$ = -1.36	-0.31 [-0.55, -0.06], k^{imp} = 3	-0.23 [-0.62, 0.15]	-0.08 [-0.34, 0.17]
11	ACPMH (2013)	**TF-CBT** vs. Combined therapies / depressive symptoms / post	*d* = -0.12 [-0.38, 0.15] (FE)	-0.12 [-0.38, 0.15], I^2 = 0 [0, 0]	7 (0)	τ = -0.05, *z* = -0.3	A = 0.37	No significant studies	-0.12 [-0.38, 0.15], k^{imp} = 0	0.08 [-1.83, 1.99]	-0.2 [0.4, 0]
12	ACPMH (2013)	**TF-CBT** vs. Combined therapies / anxiety symptoms / post	*d* = -0.09 [-0.39, 0.20] (FE)	-0.1 [-0.39, 0.2], I^2 = 0 [0, 0]	6 (0)	τ = -0.33, *z* = -0.22	A = 0.28	No significant studies	-0.1 [-0.39, 0.2], k^{imp} = 0	0.07 [-1.81, 1.94]	-0.14 [-0.41, 0.13]
13	ACPMH (2013)	TF-CBT vs. **Combined therapies** / attrition / post	MH-RR = 1.17 [0.69, 2.00] (FE)	1.1 [0.64, 1.9], I^2 = 0 [0, 0]	6 (0)	τ = -0.07, *z* = 0.7	A = 0.21	No significant studies	1.02 [0.6, 1.73], k^{imp} = 1	0.71 [0.16, 3.13]	1.07 [0.6, 1.9]
65	Imel et al. (2013)	TF-CBT vs. **Combined therapies (mostly EMDR)** / dropout / post	LOR = -0.05‡ [-0.52, 0.62] (RE)	0.05 [-0.52, 0.62], I^2 = 0 [0, 78.7]	7 (0)	τ = 0.62, *z* = 0.97	A = 0.21	No significant studies	-0.08 [-0.63, 0.46], k^{imp} = 2	-0.56 [-1.93, 0.81]	0.11 [-0.42, 0.64]
103	Powers et al. (2010)	TF-CBT vs. Combined therapies / PTSD symptoms / post	*g* = -0.07 [-0.42, 0.28] (RE)	-0.07 [-0.43, 0.29], I^2 = 48.5, [0, 91.6]	6 (1)	τ = -0.07, *z* = -0.04	A = 2.87	-0.46 [-1.3, 1.36], $L^{\hat{\mu}}$ = -0.78	-0.07 [-0.43, 0.29], k^{imp} = 0	0.23 [-2.22, 2.68]	0.06 [-0.22, 0.34]
16	Bisson et al. (2007) b	**EMDR** vs. TF-CBT/ withdrawal rate / post	MH-RR = 0.87 [0.58, 1.30] (FE)	0.87 [0.57, 1.32], I^2 = 0 [0, 0]	8 (0)	τ = 0.14, *z* = 0.87	A = 0.34	No significant studies	0.8 [0.54, 1.2], k^{imp} = 2	0.61 [0.21, 1.82]	0.82 [NaN, NaN]
27	Bisson et al. (2013)	EMDR vs. TF-CBT / leaving the study early / post	MH-RR = 1.00 [0.74, 1.35] (FE)	1.02 [0.75, 1.39], I^2 = 0 [0, 0]	8 (0)	τ = -0.21, *z* = -0.38	A = 0.23	No significant studies	1.02 [0.75, 1.39], k^{imp} = 0	1.12 [0.51, 2.46]	0.95 [0.75, 1.21]
28	Bisson et al. (2013)	**EMDR** vs. TF-CBT / PTSD symptoms (self-rated) / post	*d* = -0.30 [-0.60, 0.01] (RE)	-0.3 [-0.6, 0.01], I^2 = 32.4 [0, 90.2]	7 (1)	τ = 0.05, *z* = 0.11	A = 0.02	2.94^0 [-0.78, 18.31], $L^{\hat{\mu}}$ = 1.44	-0.3 [-0.6, 0.01], k^{imp} = 0	-0.34 [-1.39, 0.71]	-0.32 [-0.61, -0.03]
41	Chen et al. (2015)	**EMDR** vs. TF-CBT / PTSD symptoms (intrusion) / post	*d* = -0.37 [-0.68, -0.06] (FE)	-0.38 [-0.69, -0.07], I^2 = 0 [0, 0]	6 (1)	τ = -0.47, *z* = -2.19*	A = 0	-1.14 [-2.29, 0.85], $L^{\hat{\mu}}$ = -1.11	-0.21 [-0.5, 0.07], k^{imp} = 2	1.07 [-1.1, 3.25]	-0.92 [-2.24, 0.39]
42	Chen et al. (2015)	**EMDR** vs. TF-CBT / PTSD symptoms (total, sensitivity analysis with 3 studies removed) / post	*d* = -0.83 [-1.08, -0.58] (FE)	-0.84 [-1.06, -0.61], I^2 = 0 [0, 0]	8 (3)	τ = -0.07, *z* = -0.07	A = 1.46	-0.98 [-1.48, -0.66], $L^{\hat{\mu}}$ = -1.02	-0.84 [-1.06, -0.61], k^{imp} = 0	-0.8 [-1.18, -0.43]	-0.98 [-1.18, -0.77]
63	Ho et al. (2012)	**EMDR** vs. TF-CBT / PTSD symptoms / post	*g* = 0.23 [-0.03, 0.49] (FE)	0.23 [-0.03, 0.49], I^2 = 0 [0, 0]	8 (0)	τ = -0.14, *z* = -0.42	A = 0.77	No significant studies	0.23□ [-0.03, 0.49], k^{imp} = 0	0.57□ [-1.69, 2.83]	0.22 [0.01, 0.44]
82	Seidler & Wagner (2006)	**EMDR** vs. TF-CBT / depressive symptoms / post	*g* = 0.40 [0.05, 0.76] (FE)	0.39 [0.12, 0.66], I^2 = 0 [0, 0]	7 (2)	τ = -0.14, *z* = -1.14	A = 0.44	0.66 [-0.29, 1.29], $L^{\hat{\mu}}$ = -0.81	0.39 [0.12, 0.66], k^{imp} = 0	1.51 [-1.8, 4.81]	0.27 [-0.14, 0.67]
83	Seidler & Wagner (2006)	**EMDR** vs. TF-CBT / depressive symptoms / follow-up	*g* = 0.12 [-0.24, 0.48] (FE)	0.12 [-0.15, 0.4], I^2 = 0 [0, 0]	7 (0)	τ = -0.43, *z* = -1.62	A = 0.38	No significant studies	0.28 [0.02, 0.53], k^{imp} = 2	0.76 [-0.35, 1.87]	0.21 [-0.25, 0.67]

Data set No. (ID)	Author	Intervention / dependent measure / time of measurement	Original effect size [and 95% CI]	Replicated effect size, I^2 [and 95% CI]	No. of studies† (No. of sign. studies)	Begg (τ) and Egger (z)	TES	*p*-uniform [and 95% CI], pub. bias test ($L^{\hat{\mu}}$)	Trim and fill [and 95% CI], No. of studies imputed (k^{imp})	PET-PEESE [and 95% CI]	Selection model [and 95% CI]
54	Gerger et al. (2014) b	**Combined therapies** (mostly CBT) vs. Other therapies / PTSD symptoms (complex problems) / post	*g* = -0.23 [-0.42, -0.04] (RE)	-0.25 [-0.46, -0.05], I^2 = 49.1 [0, 87.5]	12 (3)	τ = -0.15, *z* = -0.55	A = 0.36	-0.63 [-1.26, -0.31], $L^{\hat{\mu}}$ = -2.29	-0.25 [-0.46, -0.05], k^{imp} = 0	-0.17 [-0.66, 0.31]	-0.22 [-0.48, 0.05]
55	Gerger et al. (2014) b	**Combined therapies** (mostly EMDR) vs. Other therapies / PTSD symptoms (non-complex problems) / post	*g* = -0.87 [-1.20, -0.53] (RE)	-0.88 [-1.21, -0.55], I^2 = 40.6 [0, 95.5]	6 (5)	τ = -0.2, *z* = -2.15*	A = 0.35	-0.85 [-1.26, -0.45], $L^{\hat{\mu}}$ = -0.01	-0.88 [-1.21, -0.55], k^{imp} = 0	0.32 [-1.19, 1.83]	-0.76 [-0.96, -0.56]
56	Gerger et al. (2014) b	**Combined therapies** (mostly TF-CBT) vs. Other therapies / PTSD symptoms (without or unclear adequate credibility) / post	*g* = -0.40 [-0.59, -0.21] (RE)	-0.42 [-0.61, -0.22], I^2 = 44.3 [0, 82.8]	15 (6)	τ = -0.03, *z* = -0.49	A = 0.66	-0.67 [-1.01, -0.35], $L^{\hat{\mu}}$ = -1.61	-0.42 [-0.61, -0.22], k^{imp} = 0	-0.31 [-0.83, 0.21]	-0.32 [-0.58, -0.06]
59	Gerger et al. (2014) b	**Combined therapies** (mostly TF-CBT) vs. Other therapies / PTSD symptoms (outliers excluded) / post	*g* = -0.43 [-0.61, -0.25] (RE)	-0.44 [-0.63, -0.26], I^2 = 45.1 [0, 82.1]	16 (7)	τ = -0.03, *z* = -0.3	A = 0.61	-0.68 [-0.95, -0.41], $L^{\hat{\mu}}$ = -1.75	-0.44 [-0.63, -0.26], k^{imp} = 0	-0.39 [-0.71, -0.07]	-0.35 [-0.62, -0.08]
60	Gerger et al. (2014) b	**Combined therapies** (mostly TF-CBT) vs. Other therapies / PTSD symptoms (outliers excluded + complex problem) / post	*g* = -0.20 [-0.36, -0.04] (RE)	-0.21 [-0.38, -0.03], I^2 = 32.5 [0, 75.1]	11 (2)	τ = 0.02, *z* = 0.36	A = 0.02	-0.48 [-0.91, -0.2], $L^{\hat{\mu}}$ = -1.85	-0.37 [-0.57, -0.17], k^{imp} = 4	-0.25 [-0.52, 0.02]	-0.21 [-0.46, 0.04]
105	Gerger et al. (2014) b	**Combined therapies** (structural equivalence no / unclear) vs. Other therapies / PTSD symptoms /post	*g* = -0.68 [-0.96, -0.40] (RE)	-0.69 [-0.97, -0.4], I^2 = 47.9 [0, 89.5]	11 (6)	τ = -0.29, *z* = -1.68	A = 0.36	-0.86 [-1.28, -0.49], $L^{\hat{\mu}}$= -1.25	-0.69 [-0.97, -0.4], k^{imp} = 0	0.08 [-0.98, 1.14]	-0.56 [-0.92, -0.19]
94	Torchalla et al. (2012)	Combined therapies (mostly other therapies) vs. Other therapies / PTSD symptoms / follow up	*g* = 0.08 [-0.03, 0.19] (RE)	0.08 [-0.03, 0.19], I^2 = 0 [0, 63.2]	8 (0)	τ = -0.29, *z* = -0.72	A = 0.6	No significant studies	0.11 [0, 0.22], k^{imp} = 2	0.1 [-0.03, 0.24]	0.08 [-0.03, 0.19]
66	Imel et al. (2013)	**Combined therapies** (mostly TF-CBT) vs. combined therapies (mostly CBT) (individual & group) / dropout / post	LOR = 0.27 [-0.34, 0.81] (RE)	0.27 [-0.28, 0.82], I^2 = 0 [0, 15.2]	9 (0)	τ = -0.11, *z* = -0.66	A = 0.47	No significant studies	0.36 [-0.17, 0.89], k^{imp} = 2	0.44 [-0.06, 0.94]	0.38 [-0.23, 1]

Note. Column 1: The treatments are grouped by treatment category and therefore the ID of the data sets is not ordered from 1 to 93. Column 3: Interventions in bold indicate the direction of the effect. Parentheses specify the interventions. If not otherwise mentioned interventions can be classified as individual therapy and face-to-face delivery. 95% confidence intervals for the I2-statistic were computed with the Q-profile method (Viechtbauer, 2007); A = Test statistic of TES; ACPMH = Australian Centre for Posttraumatic Mental Health; AMR = Applied Muscle Relaxation; BDI = Beck Depression Inventory; CAPS = Clinician-Administered PTSD Scale; CBT = Cognitive Behavioral Therapy; CI = Confidence interval; *d* = Cohen´s d; delta = Glass' Delta; EMDR = Eye Movement Desensitization and Reprocessing; Exp = Exposure; FE = Fixed-effects model; g = Hedges´ g; ITT = Intention to treat; LOR = Log odds ratio; MH-RR = Mantel-Haenszel risk ratio; OR = Odds ratio; PCT = Present Center Therapy; PE = Prolonged exposure; RE = Random-effects model with DerSimonian and Laird estimator for the between-study variance; RR = Relative risk; SIT = Stress Inoculation Training; SM = X; SN = Structured Nursing; TAU = Treatment as usual; TF-CBT = Trauma-focused Cognitive Behavior Therapy; TTP = Trauma Treatment Protocol; WL = Wait list; WMD = Weighted mean difference; w=n = Effect sizes in meta-analysis weighted by sample size. † number of studies or number of comparisons (if a study included more than one comparison). b the integration model was not specified in the paper.0 effect size estimate of p-uniform was set equal to zero (if the effect size measure was relative risk or odds ratio p-uniform's estimate was set equal to 1). c includes one to two children studies. ‡ sign in front of the original effect size appeared to be incorrect after contacting the corresponding author. □ not enough information to transform Hedges' g to Cohen's d. * $p < .05$, two-sided."