PRZEGLĄD PSYCHOLOGICZNY – THE REVIEW OF PSYCHOLOGY
2025, TOM 68, NR 4, 169–197
DOI: 10.31648/przegldpsychologiczny.12958

How to Read Meta-Analyses of Efficacy Studies and Not Go Astray.

A Primer for Psychotherapy Practitioners

Joachim Kowalski[1]
Institute of Psychology, Polish Academy of Sciences,
Experimental Psychopathology Lab
https://orcid.org/0000-0001-6281-7401

Abstract

Objective: This article aims to illustrate practical issues related to the preparation and interpretation of systematic reviews and meta-analyses of clinical trials on the effectiveness of psychotherapy. The text serves as a useful guide and map, facilitating orientation in the process of creating such studies and enabling critical evaluation of their results.

Theses: 1) The importance of systematic reviews and meta-analyses. Meta-analyses and systematic reviews are key methods of data synthesis in psychology and form the basis for evidence-based clinical decisions. In the field of psychotherapy, hundreds of review papers are published every year. **2) Variable quality of review studies.** Reviews vary in methodological quality and risk of bias, which affects the certainty of the conclusions drawn and their applicability in practice. **3) Stages of preparation and reporting.** The article describes the formal steps involved in creating systematic reviews and meta-analyses, including defining the research question, the importance of pre-registration, and the use of reporting standards. **4) Measures of effects and certainty of results.** The most commonly used measures of effect (e.g., standardised mean differences, odds ratios, number needed to treat, remission rates, or cut-off point-based indices) and assessments of certainty of results, such as measures of bias and heterogeneity, are discussed. **5) Graphical elements and additional analyses used in meta-analyses.** Graphical representations of meta-analysis results and analytical methods aimed at reducing bias are presented.

Conclusions: Systematic reviews and meta-analyses form the foundation of evidence-based clinical practice and play an important role in formulating therapeutic recommendations in psychotherapy. However, their interpretation requires awareness of the

[1] Correspondence address: jkowalski@psych.pan.pl.

processes behind their creation and the ability to critically assess the quality and limitations of these works.

Keywords: meta-analysis, systematic review, clinical trials, metapsychology

Clinical trials provide information on the safety and effectiveness of psychotherapy, limiting the influence of cognitive distortions that negatively affect decisions made in everyday clinical practice. These include cognitive biases such as illusory correlation, hindsight bias, confirmation bias, authority bias, fundamental attribution error, and many others. Clinical studies, which aim to systematically collect data using scientific methods, are partially free from some of these biases. "Partially" and "some", but it should be emphasised that they have a significant advantage over unsystematic observations in this respect. Statements such as "I don't need any research – I can see what works in my office" will not be uttered by someone who has knowledge of cognitive and social psychology, especially cognitive distortions in clinical practice (reviews in: Bowes et al., 2020 or Lilienfeld et al., 2014) or about the discrepancy between clinicians' assessments of their effectiveness and actual effectiveness (e.g., Hannan et al., 2005; Kraus et al., 2011; Walfish et al., 2012).

There are many types of clinical trials of psychological therapies[2], all of which are important for building reliable knowledge about the usefulness of therapy (APA Presidential Task Force…, 2006). In evidence-based practice (EBP), randomised controlled trials (RCTs) and their syntheses are recognised as the most important source of data on the effectiveness of interventions (e.g., Burns et al., 2011; Evans, 2003). This also applies to psychotherapy (APA Presidential Task Force, 2006; Dozois et al., 2014). In addition, the importance of RCTs with a low risk of bias is emphasised (e.g., Jobst et al., 2016), especially in the case of studies concerning isolated clinical problems (Philips & Falkenström, 2021).

However, a simple search in the PubMed database for the phrase *psychotherapy AND (trial OR case)* shows that since 2013, over 4,000 scientific papers containing these terms have been published annually. Even if we exclude the vast majority as unrelated to the topic we are interested in, we quickly realise that keeping up to date with the literature on psychotherapy research is beyond the capabilities of a single person. Thus, we face the same problem that the creators of the modern concept of meta-analysis faced (Shadish & Lecy, 2015): how to aggregate and synthesise a huge amount of data when the traditional method

[2] Definition according to Tolin et al. (2025, p. 6): "A psychological treatment is an intervention consisting of specific actions between a mental health professional and a patient or client, with the intent of engaging mental (e.g., cognitive, emotional), behavioral, or interpersonal processes, in the service of modifying health outcomes, and whose core assumptions about its procedures and mechanisms of change are founded in psychological science and are consistent with scientific understanding". In this article, due to the multitude of definitions and discussions surrounding them, when referring to psychotherapy, it should be assumed that it means psychological therapy in the above-mentioned sense.

of narrative review[3] is no longer sufficient? It is worth mentioning here that one of the first contemporary meta-analyses was created to summarise the results of several hundred studies on the effectiveness of psychotherapy (Smith & Glass, 1977).

A systematic review is a work that aims to collect, extract and synthesise data on a given topic as comprehensively as possible. A meta-analysis, or 'analysis of analyses', is a statistical aggregation of the effects of many individual studies in order to integrate their conclusions (Glass, 1976). The present work refers to both systematic reviews and meta-analyses, but for the sake of simplicity, it has been assumed in this work that the term 'meta-analysis' refers to both types of work, as a systematic review is an indispensable element of meta-analyses of independent studies.

In psychotherapy research, the integration of findings primarily concerns their effectiveness, but also other phenomena, such as adverse effects (e.g., Moritz et al., 2019) or characteristics of the therapeutic process that influence its effectiveness (e.g., Ciharova et al., 2024). In addition to aggregating large amounts of data and integrating findings from multiple studies, meta-analyses have several other advantages. One of them is a more accurate and impartial assessment of the effects obtained in studies, as researchers tend to present their results in a positive light, even if the data themselves do not necessarily justify such a positive view (Cuijpers & Cristea, 2016; Stoll et al., 2020). Another is that it allows for greater statistical power (i.e., a greater likelihood of detecting an effect that actually exists in the study) by aggregating multiple observations and narrowing the measurement error of the results (Cohn & Becker, 2003). An additional advantage is the transparency of presenting research problems and the systematic way of presenting the results of this type of work.

Nevertheless, the disadvantages of meta-analyses should also be mentioned. One of the key disadvantages is the overestimation of the effect sizes obtained in them, despite attempts to moderate the bias resulting from not including all studies (Kvarven et al., 2020). Furthermore, a search limited to titles and abstracts in the PubMed database for the phrase *psychotherapy AND (meta-analysis OR systematic review)* shows that 284 such papers were published last year alone. This means that keeping up to date with meta-analyses on psychotherapy is a difficult and time-consuming task, and this task becomes even more difficult when we consider that they vary in quality.

Therefore, this paper aims to provide a synthetic overview of the process of preparing meta-analyses and interpreting their results, taking into account effect measures, data quality assessments, and graphic elements. Another aim is to show the importance of meta-analyses in clinical practice and in the development of treatment guidelines. Due to the breadth of the topics covered, a certain degree of brevity is unavoidable. Those who would like to gain a more comprehensive understanding of the issue of research aggregation may refer to books

[3] A narrative review is a method of synthesising research in which the results presented in individual studies are described sequentially. A definite limitation of this method is that some results may be emphasised without adequate justification, e.g., based on the researcher's preferences (Higgins et al., 2024).

(e.g., Higgins et al., 2024; Rothstein et al., 2005). This work is intended to be a map and guide for psychotherapy practitioners who would like to use scientific evidence in accordance with the spirit of EBP (Spring, 2007). It may be particularly useful for those who feel that there may be a discrepancy between their professional training and their ability to understand scientific research. The aim of the work will be achieved if the reader of the meta-analysis is able to understand its purpose, the results obtained, their limitations, and the conclusions that can be drawn from all this, and at the same time, *not be led astray*.

How are Meta-Analyses Created? Research Question, Pre-Registration, and Reporting Standards

Research Question – What Do We Want to Find Out?

A properly asked question is a key element of meta-analysis, as it determines how and what kind of data we will search for and then interpret. The most popular and recommended (e.g., Tolin et al., 2015) method is to use the PICOTS acronym. A detailed explanation of the acronym is presented in Table 1 (p. 173), and below is an example of its use in the context of psychotherapy for depression. It should be added that the 'S' in this acronym can also be interpreted as study design or type, i.e., information about the type of research we are interested in (e.g., quantitative) (cf. Methley et al., 2014).

To paraphrase Chmielowski, someone unfamiliar with EBP might say that 'Everyone knows how effective psychotherapy for depression is'. However, we will rely on completely different evidence when we want to know how effective it is:

a. individual and outpatient (S) cognitive psychotherapy (I) for adolescents with a first episode of depression (P) compared to supportive psychotherapy (C) in reducing the severity of interpersonal difficulties (O) in the year following the end of therapy (T)?
b. group cognitive psychotherapy (I) conducted in a psychiatric ward (S) for people with treatment-resistant depression (P) compared to waiting for therapy (C) in reducing the severity of suicidal thoughts and frequency of attempts (O) in the period immediately after the end of treatment (T)?

The above example of two research questions, in which the only common elements are the main clinical problem and the psychotherapeutic approach, highlights the importance of precisely formulating a research question. This was already noted in the 1990s, with the practical idea that precise formulation of questions translates into accurate decisions in clinical practice (Richardson et al., 1995, as cited in Davies, 2011).

What Studies are Included in Meta-Analyses?

Meta-analyses can synthesise results from studies using different methods, e.g. a meta-analysis of correlational studies can be conducted. In studies on

Table 1

Expansion and Explanation of the PICOTS Acronym (Based on Daview, 2011 and Tolin et al., 2015)

	Meaning	Explanation
P	*Population*	The population of individuals who participated in the intervention study; e.g., people with a specific diagnosis (e.g., people with social phobia), a specific degree of severity (e.g., people after their first episode of psychosis), specific characteristics (people whose parents used physical violence), or specific life circumstances (e.g., war refugees).
I	*Intervention*	The intervention that was the subject of the study; e.g., cognitive behavioural psychotherapy, psychotherapies with an emotion regulation training component, online psychotherapies without psychotherapist support.
C	*Comparator*	Control condition against which the effects of the intervention are compared, e.g. inactive (waiting list) or active (pharmacotherapy, supportive psychotherapy).
O	*Outcome*	Evaluated effects of intervention; e.g., those related to the severity of clinical symptoms (e.g., depression, anxiety), physiology (e.g., brain activity), functioning (e.g., level of daily activity), or social and epidemiological indicators (e.g., healthcare costs).
T	*Timeline*	The duration of treatment or the period during which the effects are measured; e.g. therapy lasting 3 months or a year, measuring the effects immediately after the end of therapy or several months after its completion.
S	*Setting*	Psychotherapy setting, e.g. individual or group therapy, outpatient therapy or therapy in a psychiatric ward.

the effectiveness of psychotherapy, these may be meta-analyses of cohort studies or studies measuring change between the beginning and end of therapy (pre-test-post-test design), but most often they will be RCTs. In this research design, participants are randomly assigned to two or more conditions that differ in the type of intervention offered (or lack of intervention in the case of control conditions involving waiting for therapy or clinical monitoring). Randomisation, an element of RCTs that distinguishes this method from other clinical trials, aims to minimise differences between study groups. This makes it possible to interpret the results obtained in this way as causal effects. In this concept, individuals randomly assigned to conditions differ only in terms of their assignment, which is responsible for the differences observed between individuals. It is worth mentioning here that the pre- and post-intervention measurement scheme often used in psychotherapy research is subject to a greater risk of systematic error, as it does not allow for the control of effects related to the passage of time or being monitored by specialists, and in the case of active control conditions (discussed below), effects related to expectations or non-specific therapeutic factors.

In RCTs, at least one group should be the experimental condition, i.e. the group undergoing psychotherapy, and at least one should be the control condition. There are various control conditions in RCTs. Commonly encountered ones include treatment as usual (TAU), waiting list, and placebo therapy. TAU is difficult to define precisely, as it can vary significantly between studies (cf. Watts et al., 2015). Waiting list as a control condition involves informing patients

assigned to this group that they will have to wait for treatment. Patients in this group are examined at the beginning and end of the waiting period, which serves as a reference point for those undergoing psychotherapy during this time. Placebo therapy means treatment without 'active' elements that, according to the assumptions of the therapy, would affect the patient's well-being and health. Its effects are related only to the patient's expectations and attitude. The use of this term in psychotherapy research, where non-specific effects associated with therapeutic contact cannot be ignored (Kirsch, 2005) and where different conditions can produce different expectation effects (Boot et al., 2013), has been criticised. Another type of control condition is the use of another recognised or similarly effective therapeutic method (active control), which allows for a direct comparison of their results.

Estimates of psychotherapy effectiveness depend on the type of control condition used. In general, the effectiveness of psychotherapy compared to waiting for therapy is greater than its effectiveness compared to active control conditions (Kowalski et al., 2024; Michopoulos et al., 2021). Therefore, information about the condition to which psychotherapy is compared is crucial for understanding the synthesis of research results.

It is worth mentioning here that direct comparisons of the effects obtained in different studies are subject to serious limitations, and it should not be directly concluded that if therapy A is more effective than B (A>B) and B is more effective than C (B>C), then automatically A>C (so-called transitivity; Baker & Kramer, 2002). Studies of non-inferiority and superiority of interventions require additional factors to be taken into account, rather than direct comparison alone (Leon, 2011). There are methods of data synthesis that allow for the estimation of indirect effect sizes, e.g., network meta-analyses, but due to limited space, this topic will not be developed in this paper (those interested can refer to, for example, Caldwell et al., 2005; Hutton et al., 2015; Salanti, 2012). Nevertheless, it is simply wrong to conclude that something is more effective based on a comparison of two effect sizes obtained from the synthesis of different studies. An example of such an error would be comparing the effect of a small psilocybin RCT with meta-analyses of studies on psychotherapy and pharmacotherapy (Davis et al., 2021), where an additional level of error, apart from the naive assumption of transitivity, is the comparison of the effect of a small RCT with the effects of meta-analyses, which offer an averaged estimate but with a much narrower confidence interval.

Many RCTs also measure the stability of results in the long term after the end of therapy (follow-up). This is usually between 3 and 12 months, but follow-up measurements can be found 2 or 5 years after the end of therapy. In meta-analyses of the stability of psychotherapy effects, results from specific time intervals are usually analysed together, e.g., up to 6 months, 6–12 months, and over a year.

Meta-analyses also summarise the results of studies that did not use randomisation, and thus the possibility of drawing causal conclusions based on them is limited (e.g., Harrer et al., 2023). An example of non-random assignment to groups is the analysis of patients' symptom profiles and their assignment to

the intervention that is 'best suited' to that profile (e.g., Levi et al., 2016). Another type of non-randomised study is a cohort study with multiple measurements, in which participants are not assigned to different conditions, or this is done without the researcher's intervention.

Preregistration

In recent years, there has been a debate about the reliability and stability of results (or replicability, e.g., Nosek et al., 2022) observed in psychological research (Open Science Collaboration, 2015) and science in general (Ioannidis, 2005). One important practice that can help to obtain reliable research results is pre-registration. Its main premise is to describe the study's method (hypotheses, tools, procedures, etc.) before it begins with sufficient accuracy to shift the emphasis from exploration to hypothesis testing (Nosek et al., 2018). The aim is to avoid, for example, making hypotheses after the results are known, the effect of hindsight, or simply manipulating the parameters of statistical analyses in order to prove the existence of postulated effects at the expense of reliability (e.g. Chan et al., 2004).

In the case of meta-analyses, the importance of pre-registration is twofold. First, at the level of individual studies, pre-registration means that we can consider such a study to be less prone to bias associated with, for example, selective reporting. We can also find records in repositories indicating that such a study was being prepared. Consequently, we can take its (non)presence into account in the meta-analysis, reducing the file drawer effect (discussed below). Secondly, at the level of the meta-analysis itself, pre-registration allows for a more accurate formulation of the research question and avoids biases associated with analytical decision-making (e.g., which studies to exclude) when the results are already known. Additionally, it avoids duplication of effort among researchers, as the process of preparing a systematic review and meta-analysis can be lengthy, and the presence of pre-registration sends a clear signal to other researchers that work of this type is in progress.

Although in research on psychological and behavioural interventions (Riehm et al., 2015) and meta-analyses in psychology (Sandoval-Lentisco et al., 2025), pre-registration is a relatively recent concept and not yet a widely used practice, it has been recognised as essential in the standards for creating reviews and meta-analyses of health interventions.

Standards for Creating and Reporting Reviews

There are many guidelines for creating and reporting meta-analyses. Here, we will mention those that are publicly available and considered a standard. The purpose of such standards is to ensure the quality of meta-analyses and the clarity of their results. This is intended to result in high quality and reduced risk of bias.

The Cochrane Initiative manual (Higgins et al., 2024) plays a leading role in the creation of reviews, providing a comprehensive description of the process of preparing a synthesis of research on health interventions, including psychotherapy. In the case of standards for reporting review results, PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses, Liberati et al., 2009) is most commonly used. PRISMA is a comprehensive list of elements that should be included in the description of a meta-analysis, from the title, through the search strategy, data extraction, risk of bias assessment, to conclusions and information about the sources of funding for the study. For example, PRISMA standards specify that a work that is a systematic review or meta-analysis should have this specified in the title, which makes it easier to find such studies by narrowing the search fields in the database.

How to Interpret Meta-Analysis Results? Effect Measurements, Risk of Bias Assessment, and Data Visualisation

Effect Measurements

The most commonly used effect measures in meta-analyses are standardised mean differences (SMD), number needed to treat (NNT), odds ratios (OR) estimates, and the proportion of people who have achieved remission or significant improvement.

The latter is most often calculated in three ways. Each of them has its limitations, which are worth mentioning. The first is the reliable change index (RCI; Jacobson & Truax, 1991), in which the most important parameters are the variance of the tool's results and its temporal stability. Its assumption is to calculate a cut-off point for which there is significant certainty that the observed change can be attributed to the effect of the intervention rather than to the measurement error of the tool. There are many points of criticism regarding this index (McAleavey, 2024), including, for example, the difficulty of demonstrating significant change (i.e., conservatism), relating group variance to the results of a single patient, or loss of information through dichotomisation of a continuous variable. The second method is to assess improvement based on a cut-off point set for a given tool. This method assumes that individuals who score below a specified cut-off point have a non-clinical severity of a given symptom. It shares some limitations with RCI, such as the dichotomisation of a continuous variable. A more specific limitation is that it does not take into account the reliability of the tool, which can lead to situations where a minimal change in score, e.g., from one point above the cut-off point to one point below, leads to the conclusion that there has been a significant improvement in functioning (e.g., Andersson et al., 2025). The last method is to assess remission based on not meeting the diagnostic criteria after the end of therapy. This method, like the previous one, may result in an assessment in which a transition from relatively low symptom severity to a 'subclinical' level is treated as a significant improvement, despite not being reflected in the patient's actual level of functioning. It is

also worth mentioning the indicators from the Minimal Important Difference family, which are sometimes used in psychotherapeutic research (e.g., Copay et al., 2007).

Standardised Mean Difference (SMD)

SMDs are indicators in which the difference between two measurements (e.g., between two groups or before and after therapy within one group) is calculated and divided by the summed standard deviation (Lakens, 2013). The most popular measure of effect size used in meta-analyses is Hedges' g (which is a modified Cohen's d with an adjustment for small sample size, more in Hedges, 2025).

SMDs in intervention studies are most often in the range of –1 to 0[4], which depends, for example, on the control condition used or the type of disorder being treated with psychotherapy (e.g., Cuijpers et al., 2025), but they can also fall outside this range because they are dimensionless (Hedges, 2025). Traditionally (following Cohen), effect sizes from different ranges are interpreted as small ($d = –0.2$), medium ($d = –0.5$), and large ($d = –0.8$), but this is an arbitrary division, and there is no justification for adhering strictly to these interpretations (Lakens, 2013).

A more intuitive way of understanding SMD indicators may be to analyse the overlap between the distributions of results for two groups, e.g., those participating in therapy and those in the control group. Figure 1 (p. 178) shows two distributions of variables with effect sizes $d = 0.80$ and $d = 0.20$. For the 'large' effect, the distributions overlap by about 69%, and about 79% of those participating in therapy have scores above the control group's average score. For a 'small' effect, the distributions overlap by 92%, and approximately 58% of the therapy group has scores above the control group's average score. However, it should be noted that this interpretation is only possible if the assumptions of normality of distribution and equality of group variances are met (Hedges, 2025). Readers can check for themselves how Cohen's d can be interpreted using this application: https://rpsychologist.com/cohend/ (Magnusson, 2025).

Number Needed to Treat (NNT)

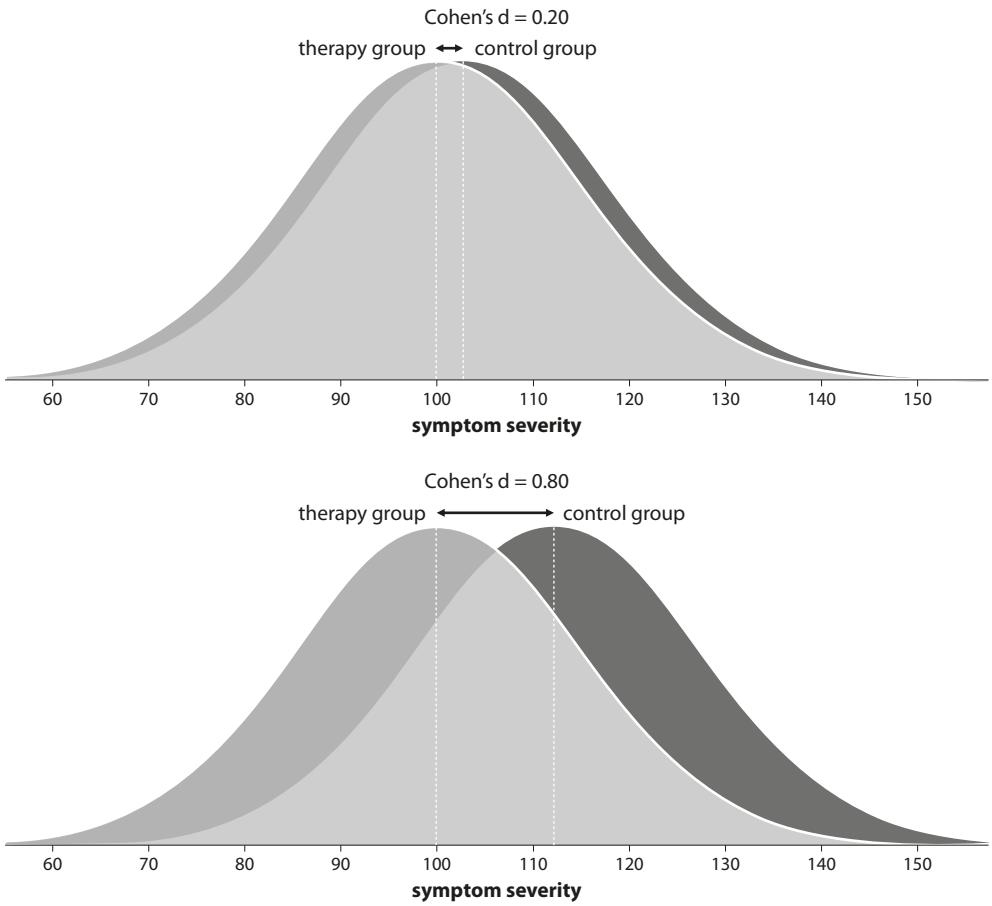
Another measurement of effectiveness is the estimated number of people who should undergo psychotherapy, compared to a control intervention, in order to achieve the desired therapeutic effect in one person. This effect may be the avoidance of an adverse event (e.g., suicide attempt) or the achievement of satisfactory improvement or remission (Hodgson et al., 2011).

The NNT is calculated by dividing the ratio of the difference in outcomes (improvement vs. no improvement, etc.) in the control group and the experimental group by the ratio of outcomes in the control group minus the ratio of

[4] Or from 0 to 1, depending on which of the two groups is the reference group (experimental or control). This means that when interpreting the effect sizes, attention should be paid to the description provided alongside them, and not just the sign, in order to correctly interpret the direction of the relationship between the groups.

Figure 1

Visualisation of two effect sizes, Cohen's d = 0.2 (small effect) and 0.8 (large effect), assuming that the results of both groups are normally distributed, their distributions were identical before the start of therapy, and the combined standard deviation is SD = 15. Prepared using a programme developed by Magnusson (2025).



outcomes in the experimental group, according to the formula NNT = 1 / (ratio of outcomes in the control group – ratio of outcomes in the experimental group). Thus, in a situation where significant improvement is observed in all individuals in the experimental group and in no individuals in the control group, the NNT is 1. If improvement is observed in half of the individuals in the experimental group and in no individuals in the control group, then NNT = 2. If improvement is observed in half of the study group and in 25% of the control group, then NNT = 4.

Unfortunately, there is no way to accurately estimate NNT based on effect size without knowing the ratios described here (cf. Furukawa & Leucht, 2011),

but Furukawa (1999) proposed a method for estimating NNT based on Cohen's d, which is important for practitioners for whom NNT may be a more intuitive indicator of treatment effectiveness than the standardised difference of means. A calculator based on this method is included in the application mentioned in the previous subsection

A complementary indicator to NNT is Number Needed to Harm, i.e., the quotient calculated in the same way, but for adverse events. This type of indicator allows for the assessment of the safety profile of psychotherapeutic intervention.

Relative Risk and Odds Ratio

Other measures used to present meta-analysis results are relative risk (RR) and odds ratio (OR). They are particularly useful for presenting dichotomous treatment effects, e.g., remission vs. no remission, as a ratio of probabilities or odds[5]. Relative risk is the ratio of the probabilities of events occurring in two groups in relation to observable effects, and the odds ratio is the ratio of the odds of an event occurring to the odds of it not occurring in two groups (e.g., Andrade, 2015; Schmidt & Kohlmann, 2008). In a situation where a denotes the achievement of a therapeutic effect in the experimental group, b denotes the absence of this effect in the experimental group, c denotes the occurrence of the effect in the control group, and d denotes the absence of the effect in the control group, the relative risk is calculated according to the formula $RR = (a / [a+b]) / (c / [c+d])$, and the odds ratio as $OR = (a/b) / (c/d)$.

OR is a more common estimate in psychotherapy research, but due to the significant difference between the two indicators when the frequency of observed events is high (Schmidt & Kohlmann, 2008) and the less intuitive interpretation of OR, it is worth illustrating the difference between the two indicators with examples.

- Example 1: 80% of patients in the therapy group achieve remission, compared to 50% in the control group. In this case, the relative risk is RR = 1.6, which can be interpreted as a 60% higher risk of achieving remission for those participating in therapy compared to the control group, and OR = 4, which can be interpreted as a four times greater chance of achieving remission among those participating in therapy (the odds in this group are 4 to 1) compared to the odds of those in the control group (odds 1 to 1).
- Example 2: If 80% of people in the treatment group achieve remission and only 20% in the control group, then RR = 4, which means a four times higher risk of remission in people in the treatment group compared to the

[5] An intuitive example that helps to understand the difference between the concepts of risk and chance is a dice. Risk is the probability that a given event will occur, i.e., the risk of rolling a 6 on a dice is $1/6 \approx 0.17$, while chance is the ratio of the probability of an event occurring to the probability of it not occurring, i.e., the chance of rolling a 6 on a dice is 1 to 5, or 0.20. (Andrade, 2015).

control group, and OR = 16, which means a 16 times greater chance of achieving remission in people in the treatment group (odds ratio 4 to 1) compared to the odds in the control group (odds ratio 1 to 4).

Confidence Intervals and Prediction Intervals

For averaged indicators such as SMD or OR in meta-analyses, confidence intervals (CI) and prediction intervals (PI) are also presented, most often for the 95% level. In general, the narrower the interval, the more precisely we are able to determine the estimated effect.

The confidence interval means – in the simplest example – that with 100 identical studies with calculated CIs, 95 of these intervals contain the true value of the effect in the population (which means that there is a risk that the interval calculated in 'our' study is one of the 5% that do not contain the population parameter). The CI is therefore interpreted as the range within which the 'true' effect size in the study population lies, the estimated certainty of the effect. CIs that touch 0 in the case of SMD or 1 in the case of OR are interpreted as statistically insignificant. This interpretation is based on the possibility that the actual effect in the population is nonexistent.

The prediction interval means – again, using the simplest example – the values of the effect sizes that we can obtain in 95 out of 100 studies similar to those included in the meta-analysis, which we could conduct in the future (IntHout et al., 2016). The PI is therefore interpreted as the range within which the effect size in the next study from the population of studies included in the meta-analysis is likely to fall, i.e., the estimated variability of the effect between studies. PIs that touch 0 in the case of SMD or 1 in the case of OR are interpreted as an indicator of high variability in the analysed results.

For example, a meta-analysis result in which the 95% CI for SMD = 0.60 is 0.20 to 1.00; but the 95% PI is –0.10 to 1.30, means that the effect of the intervention is statistically significant, but there is high variability between individual studies, which may result in the next study in the series having zero or the opposite effect to the average effect of previous studies.

Subgroup Analysis and Metaregression

Analysis of meta-analysis results broken down into subgroups allows us to determine whether any characteristics of the studies modulate the effects obtained in the study (Deeks et al., 2023). For example, in a meta-analysis that aggregates many different studies, we can expect varied therapeutic effects depending on whether whether the effect is calculated based on studies with a high or low risk of bias, individual or group therapy studies, studies involving patients diagnosed with depression or anxiety disorders, or even on which continent the study was conducted (all these analyses and others can be found, for example, in Andersson et al., 2025).

It is important that the type and number of subgroups analysed are limited, theoretically justified, and defined before the analysis is conducted (and preferably

pre-registered), and that the number of studies included in the meta-analysis is sufficient for this type of analysis (Deeks et al., 2023). An example illustrating the carelessness of subgroup analysis can be found in a seemingly humorous example from medical research, where patients were divided according to their zodiac sign, and it was shown that the position of the sun against the star systems at the time of birth is associated with a reduced protective effect of aspirin in the prevention of heart attacks (Sleight, 2000).

Metaregression is an extension of the idea of subgroup analysis (Deeks et al., 2023), except that, instead of a single categorical variable, multiple categorical and continuous variables can be analysed at once (whether there are enough studies available to perform such an analysis is a separate issue). Similar to linear regression, study characteristics serve as predictor variables, and the predicted variable is the effect size (SMD or OR).

Conducting a meta-regression allows us to address the problem of high diversity of studies included in the meta-analysis (more on this below), thanks to the possibility of identifying which study characteristics are associated with the observed heterogeneity and may modulate the observed effects (Baker et al., 2009). An example of this is a meta-analysis of 15 approaches to psychotherapy for depression, in which the estimated effectiveness of different methods varied depending on the risk of bias in the studies analysed, the type of control group, or whether the study was conducted in a Western country or not (Cuijpers et al., 2020).

How do We Assess the Quality of Meta-Analyses? Analyses of Risk of Bias and Heterogeneity

Psychotherapy studies can vary significantly, among other things, due to the populations studied, therapeutic methods, duration of treatment, setting in which therapy is conducted, etc., but also due to the quality of the research methods used. Therefore, an important element of meta-analysis is also an analysis of how this diversity may affect the interpretation of the results obtained.

Bias in clinical studies and their meta-analyses refers to the occurrence of a systematic error that may distort the results obtained. However, bias should be distinguished from non-systematic measurement error. An example of such an error is the variability observed in questionnaire measurements, which are inherently inaccurate. Bias, on the other hand, is a type of systematic error where the estimated effect sizes will deviate in a similar, systematic way from the real effect. An example of such bias is the effect of researchers' expectations, which, if it actually distorts the results, will result in the effectiveness of a given intervention being 'inflated' in seemingly independent studies.

Risk of Bias Assessment in a Meta-Analysis

When evaluating studies, we talk about the risk of bias (Boutron et al., 2023) because we do not have information about whether a given source of bias could actually have manifested itself in a given study. For example, the lack of blinding

of researchers does not automatically mean that the results are biased, but the risk of bias is higher compared to a similar study in which they were blinded. To illustrate this point, a new meta-analysis of studies on psychotherapy for depression found no significant differences between the assessment of symptom severity by participants in psychotherapy and that by blinded clinicians (Miguel et al., 2025). This may mean that, contrary to intuition, patient assessments are not more biased (e.g., due to the expectation effect) than clinician assessments.

The most commonly used tool for assessing the risk of bias in randomised trials included in the meta-analysis is the Cochrane tool for assessing risk of bias in randomised trials, currently in its updated version, i.e. RoB 2 (Sterne et al., 2019), and in the case of non-randomised trials, ROBINS-I (Sterne et al., 2016). These tools are used to assess, for example, the risk of bias resulting from: inadequate randomisation (or inadequate description of the procedure), lack of blinding, selective presentation of data on study participants or measurement tools (Higgins et al., 2023; Sterne et al., 2023). It is also worth considering the risk of bias associated with conflicts of interest, allegiance bias (Cuijpers & Cristea, 2016; Munder et al., 2013), or sources of research funding. Recently, these sources of bias have been particularly discussed in the context of research on the effectiveness of psychedelics (Buchman & Rosenbaum, 2024; Lemarchand et al., 2024). Researchers also try to take into account which studies have not been made available to the wider public (publication bias or file drawer effect). Finally, it is possible to assess how the meta-analysis itself was constructed (Boutron et al., 2023), as the decisions made by its authors may also result in a greater risk of bias.

Assessment and Analysis of the Risk of Bias of the Meta-Analysis Itself

There are several tools available for assessing systematic reviews and meta-analyses in terms of risk of bias (e.g., ROBIS, Whiting et al., 2016, the system proposed by JBI, Aromataris et al., 2015, or the ROB-ME tool for assessing bias resulting from missing studies, Page et al., 2023), but by far the most popular tool is AMSTAR (Shea et al., 2007a), currently in its second version (Shea et al., 2017), i.e., A MeaSurement Tool to Assess systematic Reviews. AMSTAR-2 allows for the assessment of both RCT meta-analyses and non-randomised studies. This tool takes into account many aspects of reviews and meta-analyses, from the formulation of the research question and pre-registration, through the selection of studies and the assessment of their bias, to the assessment of potential conflicts of interest of the authors of a given review. Studies on the first version of AMSTAR showed satisfactory inter-rater reliability, internal consistency, and ease of assessment (Shea et al., 2007b, 2009).

Quantifying the risk of bias in individual studies is difficult, so they are primarily assessed qualitatively or additional analyses are introduced (so-called sensitivity analyses, i.e. analyses showing how sensitive the result is to changes in parameters), e.g. the subgroup analyses described above. In the case of sources of bias related to missing publications (including publication bias, Sedgwick, 2015), methods such as funnel plot analysis (Egger et al., 1997, but also criticised in Lau et al., 2006), Egger's test (Egger et al., 1997), trim-and-fill analysis,

i.e. a correction for studies missing from the analysis (Duval & Tweedie, 2000), or fail-safe N estimation (Rosenthal, 1979), i.e. the number of studies with zero effect that would make the effect observed in the meta-analysis insignificant (this can be thought of as a 'testing the size of a filedrawer').

Assessment and Analysis of Heterogeneity in a Meta-Analysis

Statistical analysis of heterogeneity is based on tests that examine the extent to which differences between individual studies are due to factors other than random, unsystematic error (Deeks et al., 2023). Typically, such analyses present either Cochran's Q test or the I^2 test, which is calculated based on Q and the number of studies (Higgins et al., 2003). The result of the I^2 test is an estimate of the proportion of the variability in the results that is due to heterogeneity rather than random error, expressed as a percentage (Deeks et al., 2023). Although rarely reported in this way, the I^2 test should be interpreted together with, for example, a ence interval, as a very wide interval may make it impossible to interpret the test meaningfully (Borenstein et al., 2017). For example, $I^2 = 15\%$ can be interpreted as low heterogeneity, but a 95% CI for this I^2 estimate of 0% to 90% means that the actual heterogeneity lies in a range from very low to very high. Another indicator, perhaps even more intuitive, is the prediction interval described above (IntHout et al., 2016).

Heterogeneity in psychotherapy research (and health interventions in general) is unavoidable (Higgins & Thompson, 2002). But it also results from how the meta-analysis is designed, e.g., the broad research question or the inclusion of diverse measures of effectiveness in the analysis. Based on an example, a meta-analysis describing a small number of studies of diverse psychotherapy methods (e.g., in different settings and different therapeutic approaches) on a diverse clinical problem (which basically applies to most clinical diagnoses) will provide a 'general' answer about the effectiveness of these psychotherapies, but at the same time, the diversity of these studies will mean that the certainty of the results obtained will be significantly limited. In view of the high heterogeneity observed, researchers should decide to change their analytical strategy, e.g. not to perform a meta-analysis at all, to identify and exclude studies that are a source of heterogeneity (outliers), identify diverse effects using meta-regression, or conduct a meta-analysis based on a random effects model. More information on this topic can be found in the relevant chapter of the Cochrane handbook (Deeks et al., 2023).

How to Read All This to Understand It? Visualisation of Results and a Practical Example

Forest Plot

A forest plot is a graphical representation of meta-analysis results (Figures 2a and 2c, p. 185). It presents the studies included and the overall result of the meta-analysis in the form of successive rows, where information should be provided on: a) which study a given point represents, b) the effect obtained in

that study, together with c) the CI, most often 95%, d) in both graphic and text form. Subsequent studies are presented as points with CI 'whiskers', while the overall effect of the meta-analysis is most often presented in the form of a diamond, the vertices of which define the confidence interval limits for this analysis. Additionally, other information is included in such a graph, such as the assessment of the risk of bias, sample size, the weight of a given study for the overall effect of the meta-analysis (most often represented as the size of the point representing the effect of a given study), heterogeneity analysis (often, unfortunately, without a confidence interval), etc. The horizontal axis shows the effect sizes (SMD, OR, etc.) with an indication of which effect favours the experimental group and which favours the control group, and a marked zero point (Andrade, 2020; Sedgwick, 2015a). Visual interpretation of the graph with confidence intervals also allows for the assessment of statistical significance, as we can conclude that two independent effects differ statistically significantly ($p \leq .05$) when their CIs overlap by less than half of the average interval between them, and $p \leq .01$ when they do not overlap at all (Cumming & Finch, 2005)[6].

Funnel Plot

A funnel plot (Figures 2b and 2d, p. 185) shows a scatter plot where the points correspond to individual studies. The X-axis shows the effect sizes, the Y-axis usually shows the standard error (Sterne & Egger, 2001), and the slanted lines correspond to the confidence interval limits for specific standard error values. The purpose of this plot is to assess whether the meta-analysis shows an increased risk of bias resulting from 1. selection/ reporting bias, i.e. various circumstances resulting in a study not being included in the meta-analysis (Sedgwick, 2015b), and 2. overestimation of effect sizes by studies with small samples and/or high risk of bias (Sedgwick & Marston, 2015). We do not conclude that there is an increased risk of bias when the points are scattered symmetrically around the axis determined by the mean effect size and the scatter narrows with decreasing standard error, i.e., studies with smaller samples produce more diverse effects. The risk of bias is assessed based on the asymmetry of the funnel plot, i.e., the more 'skewed' the spread, the greater the risk of bias. Furthermore, there are statistical methods that allow for a formal analysis of the risk of missing studies or for 'filling in' the funnel plot, e.g., the trim-and-fill method mentioned above. This is particularly important as assessing funnel plots 'by eye' does not give reliable results (e.g., Simmonds, 2015).

A Practical Example

An example illustrating forest and funnel charts is a visualisation of a meta-analysis comparing the effectiveness of cognitive behavioural therapy (CBT) and

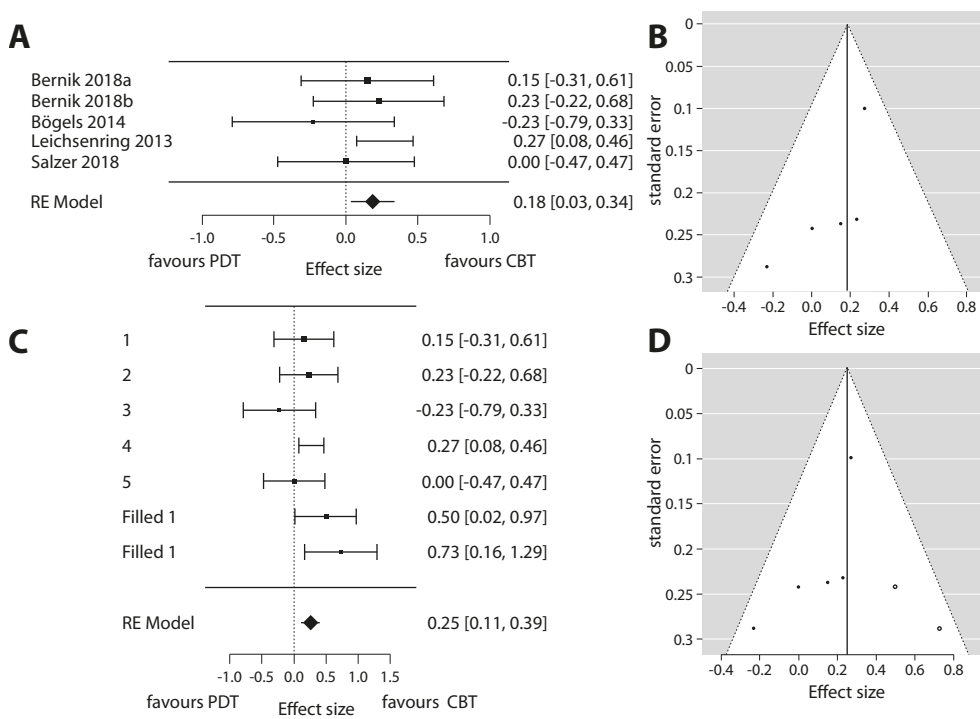
[6] This interpretation is sufficiently accurate when both groups being compared have more than 10 observations and the ratio of the wider CI to the narrower CI does not exceed 2 (Cumming & Finch, 2005).

psychodynamic therapy (PDT) in the treatment of people with social phobia, as described in Kowalski (2024). The analysis was performed using JASP v0.18.3 (JASP Team, 2024).

Figure 2a summarises the results of five studies that compared the two therapies. All but one study showed statistically insignificant differences between the two therapies (confidence intervals exceeded zero), and only one, but with the largest sample, showed a significant difference in favour of CBT. The overall meta-analysis result is a small effect in favour of CBT, SMD = 0.18; 95% CI 0.03–0.34, but with a prediction interval exceeding zero: 95% PI –0.06 to 0.42, and at the same time we observe I^2 = 3% (95% CI 0–87%) in this meta-analysis. Thus, although the I^2 estimate is only 3%, the very wide confidence interval (0–87%) means that we cannot rule out significant heterogeneity, most likely resulting from the small number of studies. In addition, the ranked correlation test for funnel plot asymmetry (Fig. 2b; Begg and Mazumdar, 1994) showed a significant (τ = –1.0, p = .017) risk of asymmetry, indicating that smaller studies in this analysis are biased towards showing zero or favouring the PDT effect (e.g. due to the investigator allegiance effect, Munder et al., 2013).

Figure 2

Forest (A) and funnel (B) plots for a meta-analysis comparing the effectiveness of cognitive-behavioural and psychodynamic psychotherapy in the treatment of social phobia. And forest (C) and funnel (D) plots after applying the trim-and-fill procedure.



Therefore, a trim-and-fill analysis was performed, which supplemented the meta-analysis with two hypothetical studies where CBT was more effective. This correction for funnel plot bias changed the overall meta-analysis result to SMD = 0.25 (95% CI 0.11–0.39), as detailed in Figures 2c and 2d (p. 185). The result of this meta-analysis itself can be interpreted as indicating a higher effectiveness of CBT in the treatment of social phobia compared to PDT, but with a small effect size.

It should be noted that, apart from the increased risk of bias demonstrated in the funnel plot analysis, the presented meta-analysis has other problems – primarily the small number of studies included and, most likely as a result of this, high heterogeneity and a wide prediction interval. This example should therefore be treated as illustrative, showing what information can be obtained by analysing forest and funnel plots.

Does This Have Any Practical Application? The Role of Meta-Analyses in Clinical Practice and the Development of Treatment Recommendations

Works synthesising research findings are crucial for EBP and clinical practice in general (e.g. Spring, 2007). The phenomenon of the gap between science and practice (e.g. Baker et al., 2008) or the cognitive distortions described in the introduction, which mean that psychotherapy practitioners do not rely on scientific evidence in their work to the same extent as, for example, intuition (Gyani et al., 2014). Therefore, the arguments in favour of the benefits of knowing and applying research findings in practice will be presented below.

One of the key arguments may be the achievement of better therapeutic outcomes. For example, in the IAPT (Improving Access to Psychological Therapies) study, cohort data from British mental health centres showed that people with severe anxiety symptoms who received the recommended therapy were more likely to improve than those who participated in therapy not recommended for this problem (Gyani et al., 2013). In practice, relying on a synthesis of data on the effectiveness of psychotherapy for a given problem allows for a more accurate selection of therapeutic methods for a given patient's problem (e.g. Kowalski et al., 2023) and adjustment of the length, intensity or setting of the treatment. It also allows for 'tailor-made therapy', i.e. adaptation to the specific mechanisms of a given problem in a given patient (e.g. Nye et al., 2023), and monitoring the effectiveness of this tailoring. In addition, the use of evidence in clinical work can have a positive effect on the ethical side of practice, as providing information and psychoeducation about the effectiveness of interventions allows for a greater degree of informed consent (Blease et al., 2018).

Another argument is the synergy with the requirement for continuous education. Drawing on scientific evidence in education and supervision can improve therapeutic work (Mallard-Swanson et al., 2021; Seegan et al., 2023). The scientific literature offers answers to questions that concern psychotherapy practitioners and can therefore be a source of support and resources for decision-making. Referring to examples from the author's experience, he was asked about the

use of prolonged exposure therapy in a person with a history of psychosis and concerns about the adverse effects of this method on symptoms other than those related to post-traumatic disorders. Meanwhile, available reviews of scientific evidence, even if they do not offer a definitive answer, indicate the safety of prolonged exposure in this patient population (Deleuran et al., 2024; Grubaugh et al., 2017), and the procedure itself does not exacerbate, but probably reduces, the severity of other psychopathological symptoms (van Minnen et al., 2015). One area of common ground for psychotherapists and researchers may be the formulation of precisely this type of question, which is relevant to practice.

Another important way of using the synthesis of scientific evidence is through therapeutic recommendations and guidelines (cf. Kowalski et al., 2024), which are formulated by professional associations or government agencies. Recommendations are formulated based on available evidence, which is then assessed in terms of the certainty with which it can be relied upon (i.e., its quality). Here, it is important to assess the risk of bias, but also inconsistency and lack of precision. The most commonly used tool for quality assessment is GRADE (Grading of Recommendations Assessment, Development and Evaluation; Schünemann et al., 2025), which is recommended, for example, in the assessment of evidence-based psychotherapy (Tolin et al., 2015).

Such recommendations are formulated, among others, by the American Psychological Association (APA, n.d.), the American Psychiatric Association (APA, 2020), and the British National Institute for Health and Care Excellence (NICE, 2014). For example, the APA has a committee within its structure that coordinates the process of developing therapeutic recommendations. APA members with clinical and scientific experience prepare the text of the recommendations in collaboration with the committee and advisors specialising in systematic reviews. The text is reviewed by both the association's authorities and experts who did not participate in the preparation of the recommendations. The British NICE has also codified the process of developing therapeutic recommendations, which take into account not only meticulously evaluated scientific evidence, but also an economic assessment of the proposed strategies for treating people with a given disorder. NICE consults stakeholders (e.g. patient groups or non-governmental organisations) on its recommendations and also has the option of subjecting them to external peer review.

Summary

The purpose of this text was to familiarise psychotherapy practitioners with issues relevant to the use of meta-analyses of psychotherapy research in clinical decision-making. The theoretical and procedural basis for compiling research syntheses was described. Much attention was paid to formal aspects, allowing for a better understanding of this type of research, but also greater awareness of its limitations. The multitude of topics covered and the cursory manner in which some of them have been treated means that this work does not in any way

exhaust the essence of the issue, but may rather serve as a gateway to further independent exploration of the subject. Therefore, an effort has been made to refer to the most basic, relevant, but also current texts, so that the article can serve as a guide for independent exploration of the literature on the subject. Good luck on your journey!

Acknowledgements

The author would like to express his sincere gratitude to Julia Szymańska, Weronika Browarczyk, Przemysław Łukasiewicz, and Magdalena Pietruch for their valuable comments on the first draft of this article.

The English version of this article was translated from Polish with the help of the DeepL translator.

References

Afonso, J., Ramirez-Campillo, R., Clemente, F. M., Büttner, F. C., & Andrade, R. (2024). The perils of misinterpreting and misusing "publication bias" in meta-analyses: An education review on funnel plot-based methods. *Sports Medicine*, *54*(2), 257–269. https://doi.org/10.1007/s40279-023-01927-9

American Psychiatric Association (2020). Development process for practice guidelines of the American Psychiatric Association – revised. https://www.psychiatry.org/getmedia/0b96df17-66a7-4f49-8159-d6522615f047/APA-Guideline-Development-Process.pdf

American Psychological Association (n.d.). APA clinical practice guideline development. www.apa.org/about/offices/directorates/guidelines/clinical-practice.

Andersson, E., Aspvall, K., Schettini, G., Kraepelien, M., Särnholm, J., Wergeland, G. J., & Öst, L. G. (2025). Efficacy of metacognitive interventions for psychiatric disorders: A systematic review and meta-analysis. *Cognitive Behaviour Therapy*, *54*(2), 276–302. https://doi.org/10.1080/16506073.2024.2434920

Andrade, C. (2015). Understanding relative risk, odds ratio, and related terms: As simple as it can get. *The Journal of Clinical Psychiatry*, *76*(7), 857–861. https://doi.org/10.4088/jcp.15f10150

Andrade, C. (2020). Understanding the basics of meta-analysis and how to read a forest plot: As simple as it gets. *The Journal of Clinical Psychiatry*, *81*(5), Article 20f13698. https://doi.org/10.4088/JCP.20f13698

APA Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *The American Psychologist*, *61*(4), 271–285. https://doi.org/10.1037/0003-066x.61.4.271

Aromataris, E., Fernandez, R., Godfrey, C. M., Holly, C., Khalil, H., & Tungpunkom, P. (2015). Summarizing systematic reviews: Methodological development, conduct and reporting of an umbrella review approach. *JBI Evidence Implementation*, *13*(3), 132–140. https://doi.org/10.1097/xeb.0000000000000055

Baker, S. G., & Kramer, B. S. (2002). The transitive fallacy for randomized trials: If A bests B and B bests C in separate trials, is A better than C? *BMC Medical Research Methodology*, *2*, Article 13. https://doi.org/10.1186/1471-2288-2-13

Baker, T. B., McFall, R. M., & Shoham, V. (2008). Current status and future prospects of clinical psychology: Toward a scientifically principled approach to mental and behavioral health care. *Psychological Science in the Public Interest*, *9*(2), 67–103. https://doi.org/10.1111/j.1539-6053.2009.01036.x

Baker, W. L., Michael White, C., Cappelleri, J. C., Kluger, J., Coleman, C. I., From the Health Outcomes, Policy, and Economics (HOPE) Collaborative Group. (2009). Understanding heterogeneity in meta-analysis: The role of meta-regression. *International Journal of Clinical Practice*, *63*(10), 1426–1434. https://doi.org/10.1111/j.1742-1241.2009.02168.x

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101.

Blease, C., Kelley, J. M., & Trachsel, M. (2018). Informed consent in psychotherapy: Implications of evidence-based practice. *Journal of Contemporary Psychotherapy*, *48*, 69–78. https://doi.org/10.1007/s10879-017-9372-9

Boot, W. R., Simons, D. J., Stothart, C., & Stutts, C. (2013). The pervasive problem with placebos in psychology: Why active control groups are not sufficient to rule out placebo effects. *Perspectives on Psychological Science*, *8*(4), 445–454. https://doi.org/10.1177/1745691613491271

Borenstein, M., Higgins, J. P., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

Boutron, I., Page, M. J., Higgins, J. P. T., Altman, D. G., Lundh, A., & Hróbjartsson, A. (2023). Chapter 7: Considering bias and conflicts of interest among the included studies. In Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., Welch, V. A. (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.4*. Cochrane. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-07

Bowes, S. M., Ammirati, R. J., Costello, T. H., Basterfield, C., & Lilienfeld, S. O. (2020). Cognitive biases, heuristics, and logical fallacies in clinical practice: A brief field guide for practicing clinicians and supervisors. *Professional Psychology: Research and Practice*, *51*(5), 435–445. https://doi.org/10.1037/pro0000309

Buchman, D., & Rosenbaum, D. (2024). Psychedelics in PeRil: The commercial determinants of health, financial entanglements and population health ethics. *Public Health Ethics*, *17*(1–2), 24–39. https://doi.org/10.1093/phe/phae002

Burns, P. B., Rohrich, R. J., & Chung, K. C. (2011). The levels of evidence and their role in evidence-based medicine. *Plastic and Reconstructive Surgery*, *128*(1), 305–310. https://doi.org/10.1097/prs.0b013e318219c171

Caldwell, D. M., Ades, A. E., & Higgins, J. P. T. (2005). Simultaneous comparison of multiple treatments: Combining direct and indirect evidence. *BMJ*, *331*(7521), 897–900. https://doi.org/10.1136/bmj.331.7521.897

Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, *291*(20), 2457–2465. https://doi.org/10.1001/jama.291.20.2457

Ciharova, M., Karyotaki, E., Miguel, C., Walsh, E., de Ponti, N., Amarnath, A., van Ballegooijen, W., Riper, H., Arroll, B., & Cuijpers, P. (2024). Amount and frequency of psychotherapy as predictors of treatment outcome for adult depression: A meta-regression analysis. *Journal of Affective Disorders*, *359*, 92–99. https://doi.org/10.1016/j.jad.2024.05.070

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, *8*(3), 243–253. https://doi.org/10.1037/1082-989x.8.3.243

Copay, A. G., Subach, B. R., Glassman, S. D., Polly Jr., D. W., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: A review of concepts and methods. *The Spine Journal*, *7*(5), 541–546. https://doi.org/10.1016/j.spinee.2007.01.008

Cuijpers, P., & Cristea, I. A. (2016). How to prove that your therapy is effective, even when it is not: A guideline. *Epidemiology and Psychiatric Sciences*, *25*(5), 428–435. https://doi.org/10.1017/s2045796015000864

Cuijpers, P., Harrer, M., Miguel, C., Ciharova, M., Papola, D., Basic, D., Botella, C., Cristea, I., de Ponti, N., Donker, T., Driessen, E., Franco, P., Gómez-Gómez, I., Hamblen, J., Jiménez-Orenga, N., Karyotaki, E., Keshen, A., Linardon, J., Motrico, E., Matbouriahi, M., Panagiotopoulou, O. M., Pfund, R. A., Plessen, C. Y., Riper, H., Schnurr, P. P., Sijbrandij, M., Toffolo, M. B. J., Tong, L., van Ballegooijen, W., van der Ven, E., van Straten, A., Wang, Y., & Furukawa, T. A. (2025). Cognitive behavior therapy for mental disorders in adults: A unified series of meta-analyses. *JAMA Psychiatry*, *82*(6), 563–571. https://doi.org/10.1001/jamapsychiatry.2025.0482

Cuijpers, P., Karyotaki, E., de Wit, L., & Ebert, D. D. (2020). The effects of fifteen evidence-supported therapies for adult depression: A meta-analytic review. *Psychotherapy Research*, *30*(3), 279–293. https://doi.org/10.1080/10503307.2019.1649732

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, *60*(2), 170–180. https://doi.org/10.1037/0003-066x.60.2.170

Davies, K. S. (2011). Formulating the evidence based practice question: A review of the frameworks. *Evidence Based Library and Information Practice*, *6*(2), 75–80. https://doi.org/10.18438/B8WS5N

Davis, A. K., Barrett, F. S., May, D. G., Cosimano, M. P., Sepeda, N. D., Johnson, M. W., Finan, P. H., & Griffiths, R. R. (2021). Effects of psilocybin-assisted therapy on major depressive disorder: A randomized clinical trial. *JAMA Psychiatry*, *78*(5), 481–489. https://doi.org/10.1001/jamapsychiatry.2020.3285

Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2023). Chapter 10: Analysing data and undertaking meta-analyses. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.4*. Cochrane. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-10

Deleuran, D. H. K., Skov, O., & Bo, S. (2024). Prolonged exposure for posttraumatic stress disorder in patients exhibiting psychotic symptoms: A scoping review.

Clinical Psychology & Psychotherapy, *31*(4), Article e3027. https://doi.org/10.1002/cpp.3027

Dozois, D. J. A., Mikail, S. F., Alden, L. E., Bieling, P. J., Bourgon, G., Clark, D. A., Drapeau, M., Gallson, D., Greenberg, L., Hunsley, J., & Johnston, C. (2014). The CPA Presidential Task Force on Evidence-Based Practice of Psychological Treatments. *Canadian Psychology / Psychologie canadienne*, *55*(3), 153–160. https://doi.org/10.1037/a0035767

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/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. *BMJ*, *315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

Evans, D. (2003). Hierarchy of evidence: A framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, *12*(1), 77–84. https://doi.org/10.1046/j.1365-2702.2003.00662.x

Furukawa, T. A. (1999). From effect size into number needed to treat. *The Lancet*, *353*(9165), Article 1680. https://doi.org/10.1016/s0140-6736(99)01163-0

Furukawa, T. A., & Leucht, S. (2011). How to obtain NNT from Cohen's d: Comparison of two methods. *PloS One*, *6*(4), Article e19070. https://doi.org/10.1371/journal.pone.0019070

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3–8. https://doi.org/10.3102/0013189X005010003

Grubaugh, A. L., Veronee, K., Ellis, C., Brown, W., & Knapp, R. G. (2017). Feasibility and efficacy of prolonged exposure for PTSD among individuals with a psychotic spectrum disorder. *Frontiers in Psychology*, *8*, Article 977. https://doi.org/10.3389/fpsyg.2017.00977

Gyani, A., Shafran, R., Layard, R., & Clark, D. M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy*, *51*(9), 597–606. https://doi.org/10.1016/j.brat.2013.06.004

Gyani, A., Shafran, R., Myles, P., & Rose, S. (2014). The gap between science and practice: How therapists make their clinical decisions. *Behavior Therapy*, *45*(2), 199–211. https://doi.org/10.1016/j.beth.2013.10.004

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, *61*(2), 155–163. https://doi.org/10.1002/jclp.20108

Harrer, M., Cuijpers, P., Schuurmans, L. K., Kaiser, T., Buntrock, C., van Straten, A., & Ebert, D. (2023). Evaluation of randomized controlled trials: A primer and tutorial for mental health researchers. *Trials*, *24*(1), Article 562. https://doi.org/10.1186/s13063-023-07596-3

Hedges, L. V. (2025). Effect sizes for experimental research. *British Journal of Mathematical and Statistical Psychology*. https://doi.org/10.1111/bmsp.12389

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

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557

Higgins, J. P. T., Savović, J., Page, M. J., Elbers, R. G., & Sterne, J. A. C. (n.d.). Chapter 8: Assessing risk of bias in a randomized trial. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.4*. Cochrane. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-08

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2024). *Cochrane Handbook for Systematic Reviews of Interventions version 6.5*. www.training.cochrane.org/handbook

Hutton, B., Salanti, G., Caldwell, D. M., Chaimani, A., Schmid, C. H., Cameron, C., Ioannidis, J. P., Straus, S., Thorlund, K., Jansen, J. P., Mulrow, C., Catalá-López, F., Gøtzsche, P. C., Dickersin, K., Boutron, I., Altman, D. G., & Moher, D. (2015). The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: Checklist and explanations. *Annals of Internal Medicine*, *162*(11), 777–784. https://doi.org/10.7326/m14-2385

IntHout, J., Ioannidis, J. P. A., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ Open*, *6*, Article e010247. https://doi.org/10.1136/bmjopen-2015-010247

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to denning meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychologv*, *59*(1), 12–19. https://doi.org/10.1037//0022-006x.59.1.12

JASP Team (2024). JASP (Version 0.18.3) [computer software].

Jobst, A., Brakemeier, E.-L., Buchheim, A., Caspar, F., Cuijpers, P., Ebmeier, K. P., Falkai, P., Jan van der Gaag, R., Gaebel, W., Herpertz, S., Kurimay, T., Sabaß, L., Schnell, K., Schramm, E., Torrent, C., Wasserman, D., Wiersma, J., & Padberg, F. (2016). European Psychiatric Association Guidance on psychotherapy in chronic depression across Europe. *European Psychiatry*, *33*(1), 18–36. https://doi.org/10.1016/j.eurpsy.2015.12.003

Kirsch, I. (2005). Placebo psychotherapy: Synonym or oxymoron? *Journal of Clinical Psychology*, *61*(7), 791–803. https://doi.org/10.1002/jclp.20126

Kowalski, J. (2024). Possible errors in a meta-analysis on the efficacy of psychodynamic therapy in social anxiety disorder (Qiqi Zhang et al., 2022). *Psychiatry Research*, *342*, Article 116174. https://doi.org/10.1016/j.psychres.2024.116174

Kowalski, J., Blaut, A., Dragan, M., Farley, D., Pankowski, D., Sanna, K., Śliwerski, A., & Wiśniowska, J. (2024). *Systematyczny narracyjny przegląd metaanaliz badań nad skutecznością psychoterapii poznawczo-behawioralnej i zaleceń terapeutycznych opublikowanych między 2010 a 2023*. [*Systematic narrative review of meta-analyses of cognitive-behavioural psychotherapies clinical trials and treatment guidelines published between 2010 and 2023*]. Polskie Towarzystwo Terapii Poznawczej i Behawioralnej.

Kowalski, J., Elżanowski, A., & Śliwerski, A. (2023). A review of selected psychotherapies for PTSD, their efficacy and treatment guidelines in adults. *Psychiatria Polska*, *58*(2), 315–328. https://doi.org/10.12740/pp/onlinefirst/157105

Kraus, D. R., Castonguay, L., Boswell, J. F., Nordberg, S. S., Hayes, J. A. (2011). Therapist effectiveness: Implications for accountability and patient care. *Psychotherapy Research*, *21*(3), 267–276. https://doi.org/10.1080/10503307.2011.563249

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434. https://doi.org/10.1038/s41562-019-0787-z

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863. https://doi.org/10.3389/fpsyg.2013.00863

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., Olkin, I. (2006). The case of the misleading funnel plot. *BMJ*, *333*(7568), 597–600. https://doi.org/10.1136/bmj.333.7568.597

Lemarchand, C., Chopin, R., Paul, M., Braillon, A., Cosgrove, L., Cristea, I., Fried, E. I., Turner, E. H., & Naudet, F. (2024). Fragile promise of psychedelics in psychiatry. *BMJ*, *387*, Article e080391. https://doi.org/10.1136/bmj-2024-080391

Leon, A. C. (2011). Comparative effectiveness clinical trials in psychiatry: Superiority, noninferiority, and the role of active comparators. *Journal of Clinical Psychiatry*, *72*(10), 1344–1349. https://doi.org/10.4088/jcp.10m06089whi

Levi, O., Bar-Haim, Y., Kreiss, Y., & Fruchter, E. (2016). Cognitive–behavioural therapy and psychodynamic psychotherapy in the treatment of combat-related post-traumatic stress disorder: A comparative effectiveness study. *Clinical Psychology & Psychotherapy*, *23*(4), 298–307. https://doi.org/10.1002/cpp.1969

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *Journal of Clinical Epidemiology*, *62*(10), 1–34. https://doi.org/10.1016/j.jclinepi.2009.06.006

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2014). Why ineffective psychotherapies appear to work: A taxonomy of causes of spurious therapeutic effectiveness. *Perspectives on Psychological Science*, *9*(4), 355–387. https://doi.org/10.1177/1745691614535216

Magnusson, K. (2025). *Interpreting Cohen's d effect size: An interactive visualization (Version 2.6.0)*. R Psychologist. https://rpsychologist.com/cohend

Mallard Swanson, K., Song, J., Beristianos, M., Aajmain, S., Lane, J. E. M., Landy, M. S. H., Suvak, M. K., Shields, N., Monson, C. M., & Stirman, S. W. (2021). A glimpse into the "black box": Which elements of consultation in an EBP are associated with client symptom change and therapist fidelity? *Implementation Research and Practice*, *2*. https://doi.org/10.1177/26334895211051791

McAleavey, A. A. (2024). When (not) to rely on the reliable change index: A critical appraisal and alternatives to consider in clinical psychology. *Clinical Psychology: Science and Practice*, *31*(3), 351–366. https://doi.org/10.1037/cps0000203

Methley, A. M., Campbell, S., Chew-Graham, C., McNally, R., & Cheraghi-Sohi, S. (2014). PICO, PICOS and SPIDER: A comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC Health Services Research*, *14*(1), Article 579. https://doi.org/10.1186/s12913-014-0579-0

Michopoulos, I., Furukawa, T. A., Noma, H., Kishimoto, S., Onishi, A., Ostinelli, E. G., Ciharova, M., Miguel, C., Karyotaki, E., & Cuijpers, P. (2021). Different control conditions can produce different effect estimates in psychotherapy trials for depression. *Journal of Clinical Epidemiology*, *132*, 59–70. https://doi.org/10.1016/j.jclinepi.2020.12.012

Miguel, C., Harrer, M., Karyotaki, E., Plessen, C. Y., Ciharova, M., Furukawa, T. A., Cristea, I. A., & Cuijpers, P. (2025). Self-reports vs clinician ratings of efficacies of psychotherapies for depression: A meta-analysis of randomized trials. *Epidemiology and Psychiatric Sciences*, *34*, Article e15. https://doi.org/10.1017/s2045796025000095

Moritz, S., Nestoriuc, Y., Rief, W., Klein, J. P., Jelinek, L., & Peth, J. (2019). It can't hurt, right? Adverse effects of psychotherapy in patients with depression. *European Archives of Psychiatry and Clinical Neuroscience*, *269*(5), 577–586. https://doi.org/10.1007/s00406-018-0931-1

Munder, T., Brütsch, O., Leonhart, R., Gerger, H., & Barth, J. (2013). Researcher allegiance in psychotherapy outcome research: An overview of reviews. *Clinical Psychology Review*, *33*(4), 501–511. https://doi.org/10.1016/j.cpr.2013.02.002

National Institute for Health and Care Excellence (2014). *Developing NICE guidelines: The manual*. https://www.nice.org.uk/process/pmg20/chapter/introduction

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nye, A., Delgadillo, J., & Barkham, M. (2023). Efficacy of personalized psychological interventions: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *91*(7), 389–397. https://doi.org/10.1037/ccp0000820

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Page, M. J., Sterne, J. A. C., Boutron, I., Hróbjartsson, A., Kirkham, J. J., Li, T., Lundh, A., Mayo-Wilson, E., McKenzie, J. E., Stewart, L. A., Sutton, A. J., Bero, L., Dunn, A. G., Dwan, K., Elbers, R. G., Kanukula, R., Meerpohl, J. J., Turner, E. H., & Higgins, J. P. T. (2023). ROB-ME: A tool for assessing risk of bias due to missing evidence in systematic reviews with meta-analysis. *BMJ*, *383*, Article e076754. https://doi.org/10.1136/bmj-2023-076754

Philips, B., & Falkenström, F. (2021). What research evidence is valid for psychotherapy research? *Frontiers in Psychiatry*, *11*, Article 625380. https://doi.org/10.3389/fpsyt.2020.625380

Riehm, K. E., Azar, M., & Thombs, B. D. (2015). Transparency of outcome reporting and trial registration of randomized controlled trials in top psychosomatic and behavioral health journals: A 5-year follow-up. *Journal of Psychosomatic Research*, *79*(1), 1–12. https://doi.org/10.1016/j.jpsychores.2015.04.010

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

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

Sandoval-Lentisco, A., Tortajada, M., López-Nicolás, R., López-López, J. A., Wagenmakers, E. J., Sánchez-Meca, J., & Hardwicke, T. E. (2025). Preregistration of psychology meta-analyses: A cross-sectional study of prevalence and practice. *Advances in Methods and Practices in Psychological Science*, *8*(1), Article 25152459241300113. https://doi.org/10.1177/25152459241300113

Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: Many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods*, *3*(2), 80–97. https://doi.org/10.1002/jrsm.1037

Schmidt, C. O., & Kohlmann, T. (2008). When to use the odds ratio or the relative risk? *International Journal of Public Health*, *53*(3), 165–167. https://doi.org/10.1007/s00038-008-7068-3

Schünemann, H. J., Higgins, J. P. T., Vist, G. E., Glasziou, P., Akl, E. A., Skoetz, N., & Guyatt, G. H. (2024). Chapter 14: Completing 'Summary of findings' tables and grading the certainty of the evidence. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, V. A. Welch (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions version 6.5.1*. Cochrane. https://www.cochrane.org/authors/handbooks-and-manuals/handbook/current/chapter-14

Sedgwick, P. (2015a). How to read a forest plot in a meta-analysis. *BMJ*, *351*, Article h4028. https://doi.org/10.1136/bmj.h4028

Sedgwick, P. (2015b). What is publication bias in a meta-analysis? *BMJ*, *351*, Article h4419. https://doi.org/10.1136/bmj.h4419

Sedgwick, P., & Marston, L. (2015). How to read a funnel plot in a meta-analysis. *BMJ*, *351*, Article h4718. https://doi.org/10.1136/bmj.h4718

Seegan, P. L., Miller, L., Young, A. S., Parrish, C., Cullen, B., & Reynolds, E. K. (2023). Enhancing quality of care through evidence-based practice: Training and supervision experiences. *American Journal of Psychotherapy*, *76*(3), 100–106. https://doi.org/10.1176/appi.psychotherapy.20220015

Shadish, W. R., & Lecy, J. D. (2015). The meta-analytic big bang. *Research Synthesis Methods*, *6*(3), 246–264. https://doi.org/10.1002/jrsm.1132

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., Porter, A. C., Tugwell, P., Moher, D., & Bouter, L. M. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology*, *7*, Article 10. https://doi.org/10.1186/1471-2288-7-10

Shea, B. J., Bouter, L. M., Peterson, J., Boers, M., Andersson, N., Ortiz, Z., Ramsay, T., Bai, A., Shukla, V. K., & Grimshaw, J. M. (2007). External validation of a measurement tool to

assess systematic reviews (AMSTAR). *PloS One*, *2*(12), Article e1350. https://doi.org/10.1371/journal.pone.0001350

Shea, B. J., Hamel, C., Wells, G. A., Bouter, L. M., Kristjansson, E., Grimshaw, J., Henry, D. A., & Boers, M. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology*, *62*(10), 1013–1020. https://doi.org/10.1016/j.jclinepi.2008.10.009

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*, *358*, Article j4008. https://doi.org/10.1136/bmj.j4008

Simmonds, M. (2015). Quantifying the risk of error when interpreting funnel plots. *Systematic Reviews*, *4*, Article 24. https://doi.org/10.1186/s13643-015-0004-8

Sleight, P. (2000). Debate: Subgroup analyses in clinical trials: Fun to look at-but don't believe them! *Trials*, *1*(1), 25–27. https://doi.org/10.1186/cvm-1-1-025

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

Spring, B. (2007). Evidence-based practice in clinical psychology: What it is, why it matters; what you need to know. *Journal of Clinical Psychology*, *63*(7), 611–631. https://doi.org/10.1002/jclp.20373

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*(10), 1046–1055. https://doi.org/10.1016/s0895-4356(01)00377-8

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, M. Borenstein, M. (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 73–98). Wiley. https://doi.org/10.1002/0470870168.ch5

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., Ramsay, C. R., Rothstein, H. R., Sandhu, L., Santaguida, P. L., Schünemann, H. J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J. C., Waddington, H., Waters, E., Wells, G. A., Whiting, P. F., & Higgins, J. P. T. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, Article i4919. https://doi.org/10.1136/bmj.i4919

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., McAleenan, A., Reeves, B. C., Shepperd, S., Shrier, I., Stewart, L. A., Tilling, K., White, I. R., Whiting, P. F., & Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, Article l4898. https://doi.org/10.1136/bmj.l4898

Sterne, J. A. C., Hernan, M. A., McAleenan, A., Reeves, B. C., & Higgins, J. P. T. (2023). Chapter 25: Assessing risk of bias in a non-randomized study. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, V. A. Welch (Eds.), *Cochrane*

Handbook for Systematic Reviews of Interventions version 6.4. Cochrane. www.training.cochrane.org/handbook

Stoll, M., Mancini, A., Hubenschmid, L., Dreimüller, N., König, J., Cuijpers, P., Barth, J., & Lieb, K. (2020). Discrepancies from registered protocols and spin occurred frequently in randomized psychotherapy trials—a meta-epidemiologic study. *Journal of Clinical Epidemiology*, *128*, 49–56. https://doi.org/10.1016/j.jclinepi.2020.08.013

Tolin, D. F., McKay, D., Forman, E. M., Klonsky, E. D., & Thombs, B. D. (2015). Empirically supported treatment: Recommendations for a new model. *Clinical Psychology: Science and Practice*, *22*(4), 317–338. https://doi.org/10.1037/h0101729

Tolin, D. F., Grasso, D., Boness, C. L., Beck, J. G., Keane, T. M., Leichsenring, F., Olatunji, B. O., Otto, M. W., & Weinand, J. (2025). A proposed definition of psychological treatment and its relation to empirically supported treatments. *Clinical Psychology: Science and Practice*, *32*(3), 213–225. https://doi.org/10.1037/cps0000220

van Minnen, A., Zoellner, L. A., Harned, M. S., & Mills, K. (2015). Changes in comorbid conditions after prolonged exposure for PTSD: A literature review. *Current Psychiatry Reports*, *17*(3), Article 549. https://doi.org/10.1007/s11920-015-0549-1

Walfish, S., McAlister, B., O'Donnell, P., & Lambert, M. J. (2012). An investigation of self-assessment bias in mental health providers. *Psychological Reports*, *110*(2), 639–644. https://doi.org/10.2466/02.07.17.pr0.110.2.639-644

Watts, S. E., Turnell, A., Kladnitski, N., Newby, J. M., & Andrews, G. (2015). Treatment-as-usual (TAU) is anything but usual: A meta-analysis of CBT versus TAU for anxiety and depression. *Journal of Affective Disorders*, *175*, 152–167. https://doi.org/10.1016/j.jad.2014.12.025

Whiting, P., Savović, J., Higgins, J. P., Caldwell, D. M., Reeves, B. C., Shea, B., Davies, P., Kleijnen, J., Churchill, R., & ROBIS group (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, *69*, 225–234. https://doi.org/10.1016/j.jclinepi.2015.06.005