The Dodo Bird, The Archbishop, and The Squirrel:
A Review and Meta-Analysis of Psychotherapy Outcome Equivalence

Shawn Sanders

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
Doctor of Philosophy in Clinical Psychology

School of Psychology
Faculty of Social Sciences
University of Ottawa

© Shawn Sanders, Ottawa, Canada, 2026

Abstract

The Dodo Bird Verdict (DBV)—the proposition that all psychotherapies are equally effective—remains controversial. Although meta-analyses have been the primary tool used to arbitrate disagreements about relative psychotherapy efficacy, the meta-analytic literature itself has not reached consensus as to whether psychotherapeutic treatments are differentially effective. In this thesis by articles, I explored the extent to which conceptual and methodological decisions in meta-analytic studies have impacted meta-analytic conclusions about the DBV.

The first chapter—a general introduction—frames the history and philosophical issues associated with meta-analysis of the DBV. The second chapter—a review article—examines several meta-analyses of psychotherapy outcomes in order to highlight key methodological issues in the DBV meta-analytic literature, including the inclusion of direct versus indirect comparisons; the restriction of analyses to *bona fide* treatments; the outcomes to be included and distinguished in the analysis; statistical considerations; and possible moderators for use in meta-regression analysis, including treatment class, disorder, allegiance, and methodological quality.

The third chapter—the meta-analytic paper—implements those methodological recommendations in meta-analytic tests of the DBV. Eligible studies for inclusion were English-language comparisons of psychotherapies for adults, compared directly to one another head-to-head, and published after the year 1980. To be retained, studies were required to (a) be randomized controlled trials, (b) involve direct comparisons of (c) "*bona fide"* psychotherapies, (d) evaluate therapies in populations with diagnosed psychological disorders, and (e) have no apparent confounds with medication use across the therapies being evaluated. Studies were excluded if they (a) reanalyzed a dataset that has already been included or (b) did not include information from which effect sizes can be calculated (e.g., means, sample sizes, variances).

Studies were collected via a hand-search of key journals (*Behavior Therapy, Behaviour Research and Therapy, Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology, JAMA Psychiatry, and Cognitive Therapy and Research, Psychotherapy, Psychotherapy Research*), a systematic term search of the ProQuest Dissertations and Theses Global database, and a search of the reference list of extant meta-analyses, collected via a systematic search of indexed subject-terms and title/abstracts in the PSYCInfo, Medline, and Proquest Dissertations and Theses databases. Risk of bias was assessed using the Risk of Bias 2 (RoB2) tool. Studies were meta-analyzed (1) all together, with randomly distributed signs, using a homogeneity test to assess significant differences, (2) all together, using the absolute value of the effect size to determine the magnitude of the effect of Treatment A vs. Treatment B, and (3) divided into treatment families (e.g., CBT, Psychodynamic), and compared using standard meta-analytic methods with CBT as the reference class. These three approaches revealed significant differences among individual treatments, (e.g., for the overall analysis of the absolute valued effect size between treatments, an average effect of $g = 0.27$, 95% CI [0.23, 0.30]). However, when divided into treatment families, the effect size within families (e.g., CBT vs. CBT), $g = 0.24$, 95% CI [0.19, 0.29], was not meaningfully different from the effect of one family vs. another (e.g., CBT vs. other), $g = 0.15$, 95% CI [0.06, 0.23].

The final chapter—a general conclusion—reviews important limitations of the data (e.g., overall high risk of bias) and explores the implications of these results both for public health and for the conceptual foundations of the DBV.

Keywords: Dodo bird verdict, meta-analysis, psychotherapy

Acknowledgements

It would simply have not been possible for me to complete this dissertation without a community of individuals who have supported me in more ways than I can possibly catalogue—but I will try. I owe gratitude to all of you.

To my supervisor, Dr. John Hunsley: thank you sincerely and profoundly for your support, patience, and guidance throughout the completion of this project. I am fortunate to have benefitted not only from your considerable technical skill, subject knowledge, and institutional know-how, but also from personal characteristics such as your equanimity, pragmatism, humour, and abundant clarity of purpose. There have been countless times over the years when I have been paralyzed by technical or conceptual details; reliably, you have disentangled me and oriented me to what really matters. There have been other times amid personal crisis or life adversity where you emerged with steady guidance, a plan, and striking compassion. I won't forget it, ever. Thank you for your input, your edits, and your ideas. Thank you for your patience and for abiding with me even through your retirement. Thank you for your effort and time as a secondary coder. Thank you for funding my meta-analysis training. Thank you for sending GIFs by e-mail. Thank you for keeping me moving forward.

To my family: thank you to my brother Adam, my grandparents Doreen, Elaine, and Neil, to Jill, to Sophie. Above all, thank you to my parents, who taught me to value education from as early back as my mind can stretch (soft stories at bedtime—and a computer in the basement, *Oceans* and *Dangerous Creatures* and *Franklin Reads*), and who never withheld a book from my devouring little hands, and who never blinked when years later, I set out on a long, arduous, and expensive educational journey—but supported me materially and emotionally, every single step of the way. Thank you both for your constant love. Thank you both for fighting your respective

battles against adversity; I wouldn't have made it here if you didn't, too. Thank you for always, always insisting (even against evidence, or against my own opposition) that I was the smartest, cleverest little Jewish boy in the whole wide world—so much that maybe a small part of me believed it, and resisted self-doubt, and gave me the temerity to look at an academic field and think I could have something interesting to say about it. Thank you for the *chutzpah*.

To Andrea, whose love and care have carried me through countless stressful days. I am forever grateful for your understanding as a fellow traveller in this program who appreciates exactly what the terrain can feel like. Thank you for your patience and for the regular doses of optimism. Thank you for teaching me to always be well-balanced in my well-being.

Thank you to my committee members Drs. Martin Lalumiere, Allison Ouimet, and George Tasca, for your input, critiques, and insights throughout the development of this project. Thank you to all three of you, and to my external examiner Dr. David Dozois, for the time and effort you have devoted to reading this dissertation and to assessing it.

Thank you to Rahmah Ikhlas for your sincere efforts as a secondary reliability coder.

To all my teachers and supervisors, both academic and clinical, thank you for the wisdom you have imparted—I am grateful that I get to carry it with me.

Thank you to the scholarship agencies that have contributed to funding my education: the Social Sciences and Humanities Research Council's Canada Graduate Scholarship program and the Ontario Graduate Scholarship program. At this specific moment in history, I especially do not take academic funding for granted.

Lastly, I am grateful to other graduate students who have modelled gracious self-compassion and dogged determination in their studies. Thank you for showing me how to move forward when the journey takes a windier route than expected.

Statement of Co-Authorship

This dissertation by articles includes two manuscripts: a methodological review published in *Canadian Psychology* (Sanders & Hunsley, 2018) and a meta-analytic study that is as-yet unpublished. Both were prepared in collaboration with Dr. John Hunsley. For the published methodological review, I was the primary author and Dr. Hunsley was the secondary author. For both the published and unpublished papers, our relative contributions were similar: my role was the formulation of specific research questions, methodological decision-making, planning and completing statistical analyses, interpreting results, and preparation of the manuscript. Dr Hunsley provided guidance and assistance throughout the process of creating this dissertation, but especially in determining the scope of the research project, for selecting appropriate research methods, and in editing the manuscripts. In addition, Dr. Hunsley served as a secondary coder for reliability of the title/abstract and full text searches for the inclusion of studies in the meta-analysis.

Generative AI was not used for any aspect of this dissertation.

Table of Contents

List of Tables and Figures

At last the Dodo said, "*Everybody* has won, and all must have prizes."

—Lewis Carroll, *Alice's Adventures in Wonderland*

We accord the Archbishop of Canterbury […] a primacy of honour and respect among the college of bishops in the Anglican Communion as first among equals (*primus inter pares*).

—*The Anglican Communion Covenant*

"Which party is right," I said, "depends on what you *practically mean* by 'going round' the squirrel. […] Make the distinction, and there is no occasion for any farther dispute."

—William James, *Pragmatism*

Hofstadter's law: It always takes longer than you expect, even when you take into account Hofstadter's law.

—Douglas Hofstader, *Gödel, Escher, Bach*

General Introduction

One of the more enduring controversies in psychotherapy research has been the Dodo bird verdict (DBV), the proposition that all psychotherapies are equally efficacious at treating psychological difficulties. This curiously named proposition refers to an appropriately fanciful scene from Lewis Carroll's *Alice's Adventures in Wonderland*. Alice watches as the Dodo bird adjudicates what he terms a "Caucus-race," a contest where dripping wet animals run about in every which direction with the aim of drying themselves off. The Dodo at last concludes the race with his verdict: "'*Everybody* has won, and all must have prizes'" (Carroll, 1865/2008).

This line appeared as an epigraph to a theoretical paper by Rosenzweig (1936), in which he argued that if several therapies with contradictory theories of change are all effective (i.e., if all have "won"), then the active ingredients in each treatment must be different than those posited by their respective underlying theories. This epigraph was in turn borrowed by Luborsky et al. (1975) and appeared for the first time as a verdict, used to summarize the authors' review of the literature—namely, that all psychotherapies were equally effective, that all had "won," and all deserve the "prizes" of equivalent scientific repute.

Following the article by Luborsky et al. (1975), the DBV became one of the first hypotheses to be evaluated with an emerging statistical technique—meta-analysis—in which a weighted average of effects is produced from those reported in numerous primary studies, with the aim of more precisely estimating the average effect in the population[1] (e.g., Smith & Glass, 1980). Meta-analysis has become the primary tool for evaluating relative psychotherapy efficacy, and correspondingly, there are now over 700 psychotherapy meta-analyses indexed in PubMed alone (Wampold & Imel, 2015). The results of these analyses have been highly heterogenous, as

[1] Assuming a random effects meta-analysis—see "Methods," later.

will be discussed later. Indeed, because of the conflicting results, such meta-analyses of psychotherapy equivalence have been cited in a number of debates in the literature, and have at turns been purported to support or refute conclusions about the relative significance of "evidence-based treatments" in psychology (e.g. Weisz et al., 2006), the importance of certain mechanisms of therapeutic change relative to others (e.g., Ahn & Wampold, 2001), or even whether psychotherapy works at all (Cuijpers, Karyotaki, Reijnders, & Ebert, 2019). Given its peripatetic history as a term—entering the psychological literature in the 1970s, alluding to a paper about common factors in psychotherapy from the 1930s, itself a reference to a children's book from the 1860s—it is perhaps appropriate that the DBV traces such a series of long-standing and interrelated debates in the psychological literature. In addition, given its relation to these provocative issues, it is perhaps unsurprising (if ironic) that the Dodo Bird refuses to go extinct.

Everything's Got a Moral, If Only You Can Find It: Implications of the DBV

Whether all psychotherapies are equally or differentially effective is an issue with potentially substantial practical implications. A large and consistent body of literature seems to suggest that psychotherapies are, generally speaking, effective in treating a variety of mental health difficulties (e.g., Barkham & Lambert, 2021). However, within this class of generally effective treatments, there nevertheless may be variation in efficacy from specific therapy to specific therapy—differentiating between them not necessarily as effective or ineffective, but instead, potentially prioritizing some as "firsts among equals." If so, then research into relative efficacy may be critical for guiding clinicians in prioritizing their therapeutic strategies when other guiding information is limited.

Similarly, research on the DBV may illuminate whether there are specific therapy approaches that are likely to better prepare trainees to be more effective clinicians, or that may more efficiently address a wider variety of presenting problems. Accordingly, such research may be of interest to graduate programs or accrediting bodies making systematic decisions about training new clinicians. If so, then the stakes of DBV research are quite high: a failure to correctly reject the DBV could lead to a promulgation of sub-optimally trained clinicians, whereas mistakenly rejecting the DBV could lead to a therapeutic monoculture that neglects potentially effective approaches or, at worst, loses them for future generations of clinicians.

Moreover, in the domain of public health policy, resources are limited, and decision-making about which treatments are offered to patients through public plans or through insurance are often by necessity based on utilitarian considerations. Research about relative efficacy may therefore be an essential guiding light to ensure those limited resources are prioritized towards treatments that, *ceteris paribus*, are most likely to help the largest number of people with the least expense. In such a context, even "small" effects may translate to substantial outcomes. For example, using a simplified conversion formula, an effect of $d = 0.20$ in favour of Treatment A over Treatment B yields a Number Needed to Treat (NNT) of approximately 9; that is, assuming all else is equal, one person would be likely to have better relative outcomes with A for approximately every 18 people given treatment (9 treated with A and 9 people treated with B; Furukawa & Leucht, 2011; Sanders & Hunsley, 2018). On the scale of a provincial or national health care program (e.g., Ontario Structured Psychotherapy Program; NHS Talking Therapies), this would translate to tens of thousands of individuals who would experience improved outcomes if given A rather than B.[2] Consequently, as with training decisions, the stakes of DBV

[2] Even if one treats the effect as continuous rather than as binary (e.g., if the effect size of $d = 0.20$ is driven by many individuals benefitting by a relatively modest amount, rather than one individual in 18 having an improved

research for public health policy are quite high: basing public policy on a mistaken failure to reject the DBV (e.g., failure to systematically fund, insure, or direct patients to a first-line treatment) could result in suboptimal clinical outcomes across an enormous quantity of people who are suffering. Conversely, basing health policy on a mistaken rejection of the DBV (e.g., systematically underfunding effective therapies; providing insurance for some therapies but not other effective treatments; directing patients towards some treatments and away from others that may help them; intellectual narrowmindedness about effective options for treatment, research, and training) may likewise undermine the aims of an effective mental health care system. DBV research, especially meta-analytic research, may therefore be critical for navigating between the Scylla and Charybdis of these potentially costly and painful errors.

The Present Manuscript

This dissertation is an attempt to clarify the question of the DBV through a comprehensive meta-analytic evaluation of the adult treatment literature. The dissertation was conducted in two parts. In the first, the disputes in the meta-analytic literature of the DBV were reviewed, with an eye to methodological factors that could account for differences among the results. Second, with these methodological issues in mind, a meta-analytic study was undertaken, evaluating the DBV in general terms (i.e., testing whether psychotherapeutic treatments are differentially efficacious *overall*, without respect to any particular clinical problem). Moderating factors were explored to determine what, if any, correlates might be associated with heterogeneity in the summary effect size (e.g., clinical problem, treatment type, methodological factors).

---

outcome), aggregating such an effect across many thousands of individuals would still represent a substantial increase in overall welfare on certain utilitarian attitudes towards maximizing the collective good (e.g., Bentham, 1789/1961).

In order to contextualize why the sort of clarification proposed by this dissertation is necessary—and why the previously outlined approach would offer such a clarification—a brief account of the DBV controversy is required. The following section will provide an overview of the history of the DBV in the adult psychotherapy literature and will outline its relation to ancillary debates in greater detail. It will also address the disagreement present in the meta-analytic DBV literature to better situate the possible contributions of this proposed dissertation to the question of psychotherapy equivalency. Finally, the precise questions that this dissertation aimed to address will be presented.

The Dodo Bird in Psychology: A Natural History

Common and Specific Factors.

Concerns about the relative efficacy of different approaches to psychotherapy (and the methods used to determine that efficacy) might be traced at least as far back as the schisms among early psychoanalysts, who were primarily concerned about the accuracy and efficacy of approaches heterodox to their own. For example, Freud's description of Adler's theories and therapeutic approach was that it was "wrong, and… dangerous" relative to his psychoanalysis (Nunberg & Federn, 1974, pp. 168–177), and Adler's derisive description of the psychoanalytic view of the mind, as that "…of a pampered child, who feels his wishes must never be denied" (Ansbacher & Ansbacher, 1956, pp. 357–358). These analysts were primarily concerned with debating the theoretical systems underlying their psychotherapies, and clinical cases of therapeutic success were often considered sufficient to establish the theoretical merits of the underlying theories (regardless of the methodological flaws of these cases; see Goleman, 1990, for an overview). In Freud's own estimation, for example, anecdotes of successful cases of

psychoanalysis rendered its assertions "independent of experimental verification" (Rosenzweig, 1985).

This position—that treatment successes reflected the success of the underlying theoretical architecture of a given psychotherapy—was challenged in the early 20th century by Rosenzweig (1936). Noting that anecdotes of success could be found in equal measure for psychoanalysis, Christian Science, and persuasion therapy (a psychiatric treatment prefiguring a form of cognitive therapy in its methods; e.g., Dubois, 1906)—and also noting that these approaches to therapy have contradictory explanations *for* their success—Rosenzweig argued that something else must necessarily be responsible for the successes attributable to each. This alternative explanation, for Rosenzweig, was the *common factors* that each of these approaches shared. His examples of such common factors included: the coherence of the proposed theory, which might provide patients with an organizing framework for integrating their personality; the therapist's personality (e.g., charisma); emotional catharsis; reconditioning and new learning; and reappraisal. Factors such as these, he argued, may be responsible for the individual successes seen by the various contradictory approaches to psychotherapy.

The link between Rosenzweig's argument and the DBV was not, at this point, empirical. Rosenzweig (1936) was not arguing for the veracity of the DBV—indeed, at this point in time, there was little empirical basis for doing so. The connection was instead theoretical—the DBV was *assumed* for the sake of argument to make the point that, if all therapies are equally effective, then the active ingredients in each psychotherapy may be other than those posited by their respective theoretical architectures. To emphasize his point about the equivalent potency of various common factors, Rosenzweig included the epigraph with the key quotation from *Alice in Wonderland*: "At last the Dodo said, 'Everybody has won, and all must have prizes'" (Carroll,

1865/2008). It was only decades later that Luborsky et al. (1975) re-appropriated the Dodo's declaration as a *verdict* on psychotherapy outcomes (thus marking the first occurrence of the DBV as it is understood today).

The Efficacy of Psychotherapy

Meanwhile, a second debate was occurring in the psychotherapy literature—namely, about whether psychotherapy was effective at all. Whereas psychoanalytic theory had historically held total dominance over the field of psychotherapy, challenges from humanistic and (especially) behaviourist schools were beginning to emerge, bringing with them new frameworks for thinking about psychotherapy efficacy—e.g., as testable. Behaviourists (e.g., Wolpe, 1948, 1958) measured outcomes in terms of client improvement rates, whereas humanists such as Rogers (1951) analyzed audiotapes and transcripts of therapy sessions to test hypotheses about effective processes in psychotherapy. Later, manualized psychotherapies would be used to provide time-limited and replicable therapy templates, which could be used in controlled comparative trials. Although treatment manuals are frequently associated with cognitive-behaviour therapies (CBT; e.g. Beck et al., 1979), the earliest such manuals were in fact for time-limited psychodynamic therapies (e.g., Mann, 1973; Sifneos, 1972).

It was in this context that Eysenck (1952) threw down the proverbial gauntlet to proponents of (non-behavioural) psychotherapies, presenting data to suggest that, in some studies, psychoanalytic and eclectic therapies had an improvement rate that was essentially equivalent to the rate of spontaneous remission—i.e., arguing that non-behavioural interventions were ineffective. Indeed, in defending the findings of his 1952 paper, he would later go on to describe dynamic theories as a "degenerative research program" (Eysenck, 1993; see generally Lakatos, 1978). These claims were met with spirited counterargument (e.g., Strupp, 1963;

Bergin, 1971), and they perhaps may have provided the impetus for an increased empirical focus on psychotherapy efficacy (e.g., through placebo- and waitlist-controlled studies of therapy efficacy). It should be noted that Eysenck (1952) had positioned his argument about remission rates as pointing to psychotherapy being ineffective overall—however, this can be misleading for contemporary readers, as he did not consider behaviour therapy to be a form of psychotherapy. Indeed, his argument about the inefficacy of psychotherapy was ultimately not about the absolute efficacy of psychotherapy overall, but about the relative inefficacy of non-behaviour therapies relative to behaviour therapies. In other words, his position was a verdict about relative psychotherapy efficacy, prefiguring and opposing the DBV.

The Caucus-Race

As previously alluded to, the debates surrounding common factors and psychotherapy equivalence came together in the work of Luborsky et al. (1975). They undertook a wide-ranging summary of the literature of comparative psychotherapy studies, counting the studies that found a statistically significant relative effect, as well as those that did not, and quantitatively summarizing this count. The authors also notably included an assessment of the methodological quality of each study. They concluded that there was little evidence that one type of therapy was superior to another—a conclusion that they termed the DBV—and suggested that this lack of treatment differences was evidence in favour of the importance of common factors.

The analysis by Luborsky et al. (1975) was a significant first attempt to quantitatively summarize the literature on the relative efficacy of psychotherapies. However, their method of counting the number of statistically significant studies and comparing them to the number of statistically non-significant studies—a process referred to by Borenstein et al. (2009) as "vote-counting"—is not an appropriate method of inference. Although it is true that a significant

finding supports the probable existence of a true effect, it is not true that the absence of a significant finding is evidence for an absence of an effect, as low power and imprecision may be responsible for the null findings rather than a lack of a true effect. Indeed, due to low sample sizes in many areas of clinical psychology research, studies can generally be expected to fail to detect small effects four out of five times, and to fail to detect even the largest effects one out of five times (Rossi, 1990). Luborsky et al.' (1975) study therefore does not offer the compelling evidence for the DBV that its authors may have believed. Nevertheless, their study set off a flurry of attempts to assess the veracity of the DBV, and in so doing, played a role in the development of an entire new branch of statistical analysis.

Through the Looking Glass: Meta-Analyses of the DBV

Key Early Meta-Analyses

The study by Luborsky et al. (1975) invigorated the question of psychotherapy efficacy and, in particular, the question of how to best determine the consensus of the relatively recent abundance of clinical trials of psychotherapy. It was in this context that a new method was developed to summarize the effects of a given literature—namely, meta-analysis. Meta-analyses would become the key method for assessing psychotherapy efficacy, and numerous meta-analyses on the topic would be produced in the following decades. However, as the following review will demonstrate, these analyses did not produce a corresponding consensus about psychotherapy equivalence.

Smith et al., 1980. The earliest application of the modern meta-analytic method was in the field of psychotherapy research. Glass and his colleagues, building on previous statistical work (e.g., Cohchran & Carroll, 1953), first presented their meta-analytic method as a means of synthesizing the research on psychotherapy outcomes (e.g., Glass, 1976; Smith & Glass, 1977).

They then expanded this work into a substantial work of scholarship, *The Benefits of Psychotherapy* (Smith et al., 1980). Their work was designed to address arguments that reviews of the psychotherapy outcome literature had, so far, been plagued by selective readings of the data (e.g., Bergin, 1971; Rachman, 1973; Lambert, 1976). By including a wide range of studies and summarizing them statistically, the authors hoped to bypass these arguments. Later, Glass (2000) would go on to describe additional, more personal motivations for this line of research:

> "I was so impressed with the power of psychotherapy as a means of changing my life and making it better that by 1970 I was studying clinical psychology […] the weight of academic opinion at that time derived from Hans Eysenck's frequent and tendentious reviews of the psychotherapy outcome research that proclaimed psychotherapy as worthless—a mere placebo, if that. I found this conclusion personally threatening—it called into question not only the preoccupation of about a decade of my life but my scholarly judgment (and the wisdom of having dropped a fair chunk of change) as well. […] I wanted to take on Eysenck and show that he was wrong: psychotherapy does change lives and make them better."

The findings by Smith et al. (1980) seemed to suggest that, overall, psychotherapy (behavioural and otherwise) is effective in an absolute sense. The findings are also frequently taken as evidence in favour of the DBV—i.e., supporting relative equality for all approaches to psychotherapy. This conclusion of psychotherapy equivalence was met with both ovation and opprobrium.

Critiques Of the Meta-Analytic Method. A number of behaviourists (e.g., Eysenck, 1978; Rachman & Wilson, 1980) critiqued the Smith et al. (1980) results on the basis of the inappropriateness of meta-analysis altogether. They argued that these methods prioritize hyper-

inclusivity over methodological rigour, and that well-designed individual studies are superior to a meta-analytic hodgepodge. However, a problem with this type of criticism is that even well-designed studies will *necessarily* disagree with one another. A statistical power to detect an effect of 80% still entails that, even among 10 well-designed studies, researchers will fail to detect a true positive effect in two studies. The issue becomes worse if one considers the false positive rate evident across published studies (e.g., Ioannidis, 2005). Furthermore, for most studies in clinical psychology, the power to detect a true effect is substantially less than 80% (Rossi, 1990). The results of individual studies, in other words, are *expected* to be inconsistent—it is only by their synthesis that the limitations of each in terms of statistical power can be overcome (Borenstein et al., 2009). That said, the criticisms of Eysenck (1978) and Rachman and Wilson (1980) did point to key methodological considerations for meta-analytic studies of the DBV—namely, that meta-analyses can potentially be biased by methodologically poor studies; that inclusion criteria must be appropriately strict; and that large overall meta-analytic conclusions tend to average across (and thereby disguise) important variability (see also Cronbach, 1982). Fortunately, current meta-analytic methods can to some degree account for variability, methodological and otherwise, through subgroup analysis and meta-analytic regression.

Other Methodological Critiques. One of the conclusions of Smith et al. (1980) was that psychotherapies, of all stripes, are all generally equally effective. However, the authors reported on a variety of differential treatment effects—for example, that cognitive therapies (CT) and CBT generally had larger effect sizes overall than did humanistic psychotherapies, and that similar differences also existed between treatments when looking at their effects for particular disorders. In one of the many sets of analyses they reported, the authors analyzed their data by

grouping treatments into superordinate classes of "verbal" treatments (e.g., cognitive, psychodynamic) and "behavioural" treatments (e.g., behaviour modification, systematic desensitization). Statistical comparisons between these two treatment classes were nonsignificant, and it is these results that have frequently been cited to support the DBV. In light of Eysenck's (1952) assertions that non-behavioural psychotherapies were ineffective, and Glass's apparently personal motivation to overturn these claims, a grouping such as this is perhaps understandable. However, as Hunsley and DiGiulio (2002) pointed out, it's unclear what *theoretical* reason could be offered to group cognitive therapies with psychodynamic and humanistic therapies, rather than with behaviour therapies. Cognitive and behavioural therapies have long been viewed as part of the same "family" of treatments by (for instance) professional organizations such as the Association for Behavioral and Cognitive Therapies (previously the Association for Advancement of Behavior Therapy) and the National Association of Cognitive-Behavioral Therapists; in contrast, no professional organization focused on psychodynamic or humanistic therapies has ever included cognitive therapy within their purview. Indeed, Eysenck himself considered cognition to be a form of mental behaviour (Eysenck & Martin, 1987). Setting aside the results associated with this apparent taxonomic error, most meta-analytic results reported by Smith et al. (1980) do *not* support the DBV, contrary to what is commonly taken to be the case (e.g., Wampold et al., 1997; Wampold et al., 2017).

In addition to this issue, the findings of Smith et al.'s (1980) meta-analysis did not account for statistical dependencies—for example, between multiple outcomes in a single study, which were treated as statistically independent by Smith et al. (1980). Although techniques for handling dependencies were not available to Smith et al. (1980) at the time, such dependencies are nevertheless problematic for straightforwardly interpreting their results. Meta-analytic

methods generally assume that data are not correlated—if there are dependencies in the data (e.g., if two supposedly independent outcomes were actually measured in the same set of people), this could artificially inflate the calculated effect sizes (and depending where the dependencies occur, could increase some effect sizes but not others). A second issue relates to the authors' inclusion of unpublished dissertations in their analysis. Although criticized as overly inclusive and as diluting the methodological quality of the included studies (e.g., Eysenck, 1978), concerns about publication bias suggest that the inclusion of high-quality unpublished studies may in fact be valuable (e.g., Hopewell et al., 2005). Regardless, both points would be addressed in future meta-analyses and replications (see below).

Other Meta-Analyses, 1980 to 1997. Following the publication of Smith et al. (1980), a large number of meta-analyses were published examining the relative efficacy of psychotherapies. The approaches undertaken by the authors of these analyses often varied considerably. One early approach was to evaluate the data of Smith et al. (1980). Andrews and Harvey (1981), for example, undertook analyses of the subset of studies from Smith et al. (1980) that examined a clinical population. Significance tests revealed that therapy classes were significantly different from one another; relative to control conditions, verbal therapies (cognitive, gestalt, psychodynamic) had an effect of $d = 0.74$ whereas behaviour therapies (including cognitive behavioural) had an effect of $d = 0.97$. Developmental therapies (including client-centred and counselling) had an effect of $d = 0.35$, and unlike the other therapies, were not significantly different from the effect of placebo versus waitlist. No indication was provided as to what constituted a control (e.g., waitlist, placebo, treatment as usual), and as in the original Smith et al. (1980) study, statistical dependencies did not appear to be accounted for.

Other meta-analyses approached the question of psychotherapy efficacy comprehensively. Shapiro and Shapiro (1982), for example, conducted an independent and wide-ranging meta-analytic evaluation of psychotherapy outcome studies, extracting effects from studies that compared two or more treatments to each other or to a control condition within a single study (as opposed to Smith et al., 1980, who extracted effects from numerous disparate studies). This approach had the advantage of reducing certain sources of error (e.g., non-equivalencies between studies in terms of treatment setting, delivery, control condition, etc.). In addition to the overall effect versus a no-treatment condition, Shapiro and Shapiro broke down effect sizes by treatment class, target problem, and type of measure. They also separately analyzed direct comparisons of treatments (rather than relative to a control comparator), and undertook regression analyses on possible moderators of outcome, including therapy variables, client variables, contextual variables, measurement variables, and design variables. The overall effect of behavioural, cognitive, and mixed treatments relative to control conditions were found to have a significant advantage over the average overall effect of all other treatments relative to control, d = 0.32, 0.40, and 0.52 respectively. By contrast, dynamic/humanistic therapies, unclassified, and minimal treatments were found to have a significant disadvantage compared to the average effect of other treatments versus control, d = -0.53, -0.23, and -0.56, respectively. However, in direct comparisons of dynamic/humanistic therapies versus behavioural methods (i.e., social skills training), the relative advantage of behavioural therapy disappeared (although low statistical power may be implicated). Direct comparisons revealed a significant advantage of cognitive therapy to systematic desensitization, d = 0.53. Cognitive therapies had the largest effects for anxiety/depression, social/sexual problems, and performance anxieties, whereas behaviour therapies had the largest effects for phobias and physical/habit problems. However,

the effects for cognitive therapy were based on relatively few studies and it is unclear to what extent the therapies included in the comparisons were representative of therapies as undertaken in clinical practice (i.e., were *bona fide* therapies). Moreover, it's unclear what principled distinctions led the authors to categorize "performance anxieties" (which, in addition to combining public speaking and test anxieties, included academic underachievement and vocational indecisions") as a conceptually distinct problem area from "social and sexual problems" (which included class discussion difficulty, assertion, and social inadequacy in addition to dating anxiety, marital communication, and sexual difficulty).

Some meta-analyses examined the effects of psychotherapies within a single disorder. For example, Robinson et al. (1990) tested the effects of psychotherapy for major depression, examining the relative effects of four classes of treatments —cognitive, behavioural, cognitive-behavioural, and general verbal (which included psychodynamic, client-centred, and interpersonal therapies)—relative to control conditions and in direct comparisons to one another. The authors also undertook ratings of researcher allegiance and regressed the results of the comparisons on these ratings. Relative to control, the effects for cognitive, behavioural, and cognitive behavioural—d = 0.96, 1.02, and 0.85, respectively—were similar, and higher than the effects of the verbal therapies, d = 0.49. In direct comparisons, the authors found significant relative advantage of cognitive-behavioural therapies relative to behavioural therapies for depression, d = 0.24. The verbal class also fared significantly worse relative to cognitive therapy, d = 0.47, behavioural therapy, d = 0.27, and cognitive-behavioral therapy, d = 0.37. However, these relative advantages disappeared when researcher allegiance was taken into account. The inclusion of these allegiance ratings—something first employed by Smith et al. (1980)—was a methodological strength of this study, although the criteria by which the authors determined

allegiance appeared vague. Furthermore, their categorization scheme (i.e., separating cognitive and behavioural from cognitive-behavioural; categorizing heterogeneous treatments like psychodynamic and client-centred as "general verbal") was of questionable utility, and no provision was made to exclude pseudo-treatments from the comparisons.

Still other meta-analyses examined the effect of a single treatment orientation across clinical problems. Miller and Berman (1983) examined the effects of CBT relative to no treatment or other therapies. Here, CBT was distinguished as any therapy using a particular rationale (viz., examining maladaptive beliefs to address emotional difficulties), rather than as a broad umbrella category encompassing therapies based on either cognitive and/or behavioural techniques. However, the authors only distinguished one specific alternative therapy as a comparator—systematic desensitization—and classified all others into a broad "other therapy" category. They further broke their analysis down by post-treatment versus follow-up data, by clinical problem, by outcome measure, by modality (i.e., individual or group therapy), and by patient characteristic. Overall, CBT was found to be superior to "other" therapies, but not desensitization, at both post-treatment ($d = 0.21$) and follow-up ($d = 0.24$). Relative to all other therapies (including desensitization), CBT evinced greater efficacy for anxiety, $d = 0.17$, but not depression or somatic difficulties. Although the domain-specific analyses were methodologically admirable, the authors' treatment classification scheme did little to illuminate which treatments would be expected to show differential efficacy, or under which circumstances this might be expected to occur.

Other meta-analyses examining a single orientation across all problems include Svartberg and Stiles (1991) and Anderson and Lambert (1995), both of which examined short-term psychodynamic psychotherapies (STPP). Svartberg and Stiles (1991) compared STPP to control

conditions and other treatments at post-treatment, 6-month follow-up, and 12-month follow-up. Relative to other treatments, they found STPP to be significantly less efficacious, $d = 0.24$, a gap which grew at 12-month follow-up, $d = 0.72$ (no value corrected for heterogeneity of effect sizes was reported at 6-month follow-up). When breaking down by comparator treatment, STPP was found to be less efficacious than CBT post-treatment, $d = 0.47$, and at follow-up, $d = 0.45$. It was also found to be less efficacious post-treatment than experiential therapies, $d = 0.53$, but not at follow-up. There were no differences found between STPP and behavioural therapies or "nonspecific" therapies. In stark contrast to these results, Anderson and Lambert (1995)—who compared STPP to control conditions, minimal treatment conditions, and alternative treatment—found no difference between STPP and alternative treatments. This was the case at both post-treatment, $d = -0.02$, 95% CI [-0.08, 0.04], and at follow-up, $d = 0.04$, 95% CI [-0.05, 0.12]. Relative to CBT, STPP also showed no significant differential efficacy, $d = 0.01$. Anderson and Lambert attributed the inconsistency between their findings and those reported by Svartberg and Stiles to including a greater number of studies in their analysis, to having adjusted effect sizes to account for pre-treatment differences between groups, and to Svartberg and Stiles having included a number of studies which may have constituted inadequate tests of STPP (including non-*bona fide* STPP such as treatments with too few sessions and treatments that were actually long-term psychodynamic therapy or integrative therapy). Another explanation for the different findings may lie in Anderson and Lambert's classification of "alternative therapies," for which one possible inclusion criterion was being "expected to produce results similar to [STPP]" (Anderson & Lambert, 1995, p. 505). If comparator treatments were specifically selected to produce effects that did not differ from STPP, and they were potentially excluded if they were expected to differ from it, then it perhaps comes as no surprise that the authors found no

difference between STPP and comparators; however, a comparison based on such apparently circular reasoning arguably reveals little. Neither Svartberg and Stiles (1991) nor Anderson and Lambert (1995) accounted for researcher allegiance in their results, although the former study included a rating of methodological quality as a moderator.

Lastly, some meta-analyses in this period focused on one therapy within one clinical problem area. Dobson (1989), for example, evaluated cognitive therapy for depression, examining its effects relative to waitlist, pharmacotherapy, behaviour therapy, or other therapies. Cognitive therapy was found to be more efficacious than behaviour therapy, $d = 0.46$, and other therapies, $d = 0.54$. The only outcome measure assessed was the Beck Depression Inventory, and no follow-up data were reported. As has been the critique for previously reviewed meta-analyses, the constituent therapies of the "other therapy" category were heterogeneous (ranging from assertiveness training to psychodynamic therapy to "therapy as usual"), complicating interpretation of the results; moreover, no considerations were made for allegiance or methodological quality, and the criteria by which studies were determined to have included depressed participants was unclear.

Wampold et al., 1997

The meta-analysis by Wampold et al. (1997) stands out from other meta-analyses in the psychotherapy literature for several reasons. First, this study is one of the most cited in the psychotherapy outcome literature (over 2,000 times as of September 2025, according to Google Scholar), and it is the basis for much of the lead author's discussion of the DBV in his book, *The Great Psychotherapy Debate* (Wampold, 2001; Wampold & Imel, 2015). This book itself has been cited over 10,000 times as of September 2025 (exceeding the citation count of Smith & Glass, 1977) and it has widely been viewed as the "final word" on the DBV for many

psychologists. Second, Wampold et al.'s (1997) distinctive methodology deserves particular attention, as one of its methodological decisions—to examine comparisons of only *bona fide* therapies—remains highly influential among psychotherapy meta-analyses.

Wampold et al. (1997) availed themselves of the greatly improved statistical sophistication of meta-analytic techniques relative to Smith and Glass's (1977) initial analyses. In addition to employing these technical improvements, Wampold et al. (1997) also introduced greater methodological controls in order to improve the internal validity of their results. For instance, the authors limited their inclusion of studies to only direct, head-to-head comparisons of psychotherapies published in high quality journals, thereby controlling for a number of extraneous factors that can arise when treatments are compared on the basis of their effect sizes relative to control conditions (e.g., wait list controls, attentional controls). They also limited their analysis to *bona fide* psychotherapies—defined as those offered by genuine psychotherapists (i.e., individuals with at least a master's degree who made active treatment decisions), geared towards a broadly-defined clinical problem, and meeting at least two of the following criteria: citing an established psychotherapeutic approach, citing psychological principles in their rationale, being manualized or codified in a professional book, or having their active therapeutic ingredients identified and cited. This focus on both *bona fide* psychotherapies and direct comparisons, although not entirely novel in the literature (e.g., Lambert & Bergin, 1994), has become somewhat of a "gold standard" in the meta-analytic DBV literature. In particular, the *bona fide* criterion ensures a minimal level of conservativeness about the definition of psychotherapy, excluding such treatments as "rebirthing therapy" which are of questionable theoretical provenance to nearly all practicing psychologists, as well as treatments merely provided as control conditions.

Some aspects of Wampold et al.' (1997) methodology, however, were highly idiosyncratic. In their analysis, the authors chose to dispense with treatment classes, instead examining the distribution of *all* treatments' randomly-signed effect sizes around zero. That is, the effect sizes for each of the comparisons between two treatments were randomly assigned either a positive or negative sign, producing a distribution of effect sizes centred around zero (as random distribution of positive and negative signs to the study effect sizes produces a mean effect of approximately zero). The authors then examined the degree of homogeneity of the effects—tight, homogeneous clustering around zero was taken to indicate that treatment differences generally did not differ substantially from zero, whereas a large degree of heterogeneous spread was taken to indicate that treatments differed by a correspondingly large degree. This stands in contrast to the method used in most meta-analyses (indeed, it seems like almost *all* meta-analyses apart from those from Wampold's research group, or explicit replications of them; Tran & Gregor, 2016). Most meta-analyses produce a summary effect by averaging all included effects, weighted by their precision; to ensure that the valence of the effects do not cancel out, either (a) one reference class is chosen and all other effects are given a sign relative to that class (e.g., "a positive sign indicates an effect superior to CBT, whereas a negative sign indicates an effect inferior to CBT") or (b) the absolute effect size is taken, which gives no information about directionality of the effect but nevertheless provides information about the difference of the overall effect from zero (i.e., if you are trying to answer the question "overall, are psychotherapies different from one another?" then the magnitude, not the direction, of differences is all that should matter). Although Wampold et al. (1997) undertook a supplemental meta-analysis of the absolute effect size ($d = 0.21$), they did not consider this to be

the primary or correct method of analysis, and they did not derive their conclusions about the DBV from this analysis.

With their random-signs technique, the authors reported no evidence of heterogeneity, which they took as indicative of support for the DBV. However, this analytic method is potentially problematic, as most of the comparisons included in the analysis were comparisons of one type of CBT to another type of CBT (Crits-Christoph, 1997; Hunsley & DiGiulio, 2002). Homogeneity of effects may not be very informative if most of the treatment comparisons pitted conceptually similar treatments against one other. Indeed, as I will later argue, it may be that the most critical test of the DBV is undertaken by evaluating differences between conceptually *dissimilar* treatments.

The conclusion to which Wampold et al. (1997) arrived was that *bona fide* psychotherapies are equally effective, providing support for the notion that what is efficacious in psychotherapy is not techniques specific to any given psychotherapy approach, but rather, shared, and theoretically noncentral factors (i.e., the "common factors"). Elsewhere, they have similarly indicated that data in support of the DBV constitutes evidence of favour of the "common factors" argument (e.g., Wampold & Imel, 2015). This connection between the DBV and the common factors argument, harkening back to Rosenzweig, will be addressed later ("Orthogonality to common factors argument," below).

Recent History

Since 1997, the number of meta-analyses undertaken on psychotherapy outcome has grown enormously: a decade-old count found 700 psychotherapy meta-analyses indexed in PubMed alone (Wampold & Imel, 2015), a number which has certainly grown since then. Indeed, since work on this dissertation began, the METAPSY project has released a tool to

generate customizable, on-demand meta-analyses of the psychotherapy-versus-control trials contained in their open-access and regularly updated databases (e.g., Cuijpers, Miguel, et al., 2025). Comprehensively reviewing all psychotherapy outcome meta-analyses is not practically feasible; however, since Wampold et al.'s (1997) study, only a small subset of these meta-analyses has examined direct, head-to-head comparisons of psychotherapies, as opposed to indirect comparisons to control conditions (e.g., waitlist, placebo). A title/abstract of the PsycINFO database (as of June 2025) was used to produce the summary information presented in Table 1 about these meta-analyses (see Appendix A for search terms and exclusion criteria):

Table 1

Meta-Analytic Conclusions from Direct Comparisons of Psychotherapies, November 1997 to June 2025

Study	Across or within disorders	Direct comparators	Authors' conclusions support treatment equivalency/DBV
Tolin (2010)	Across	CBT vs. other therapy	No
Ruiz (2012)	Across	ACT vs. CBT	No
Marcus et al. (2014)	Across	CBT vs. ACT, behavioural, IPT, psychodynamic and "other" psychotherapies	No
Kivlighan et al. (2015)	Across	Psychodynamic vs. non-psychodynamic therapy	Yes
Goldberg et al. (2018)	Across	Mindfulness vs. other therapies	No
Munder et al. (2019)	Across	"Strengths-based" CBT/dynamic therapies vs. "deficit-based" CBT/dynamic therapies	No
Linardon et al. (2019)	Across	Dropout rates for IPT vs. other psychotherapies	No
Wampold et al. (2002)	Within	CT vs. other therapies for depression	Yes
Kotova (2005)	Within	IPT vs. other therapies across problems for, depression, and for eating disorders	Yes
Bisson et al. (2007)	Within	EMDR vs. CBT vs. stress management vs. "other"	No

Table 1

Meta-Analytic Conclusions from Direct Comparisons of Psychotherapies, November 1997 to June 2025

Study	Across or within disorders	Direct comparators	Authors' conclusions support treatment equivalency/DBV
		therapies for PTSD/complex PTSD	
Siev & Chambless (2007)	Within	CT vs. relaxation for GAD and panic disorder	No
Benish et al. (2008)	Within	Psychotherapies (undifferentiated) for PTSD/trauma	Yes
Ekers & Gilbody (2008)	Within	Behavioural therapy vs. others (cognitive, brief, and supportive therapies) for depression	No
Currier (2010)	Within	CBT vs. other therapy for grief	No
Ougrin (2011)	Within	CT vs. exposure for panic disorder, social anxiety, PTSD, and OCD	No
Ho & Lee (2012)	Within	CBT vs. EMDR for PTSD	No
Baardseth et al. (2013)	Within	CBT vs. non-CBT for anxiety	Yes
Braun et al. (2013)	Within	CBT vs. behavioural activation vs. psychodynamic vs. interpersonal vs. supportive for depression	Yes
Budge et al. (2013)	Within	"Evidence-based therapies" vs. "treatment as usual" for personality disorders; *bona fide* therapies vs. one another for personality disorders	No
Newton-Howes & Wood (2013)	Within	CT vs. supportive therapy for schizophrenia	Yes

Table 1

Meta-Analytic Conclusions from Direct Comparisons of Psychotherapies, November 1997 to June 2025

Study	Across or within disorders	Direct comparators	Authors' conclusions support treatment equivalency/DBV
Spielmans et al. (2013)	Within	CBT vs. non-CBT for bulimia nervosa and binge eating disorder	No
Cuijpers, Karyotaki, Pot, Park, & Reynolds (2014)	Within	CBT, problem-solving therapy, behavioural activation, psychodynamic, and supportive therapy vs. one another for geriatric depression	No
Cuijpers et al. (2014)	Within	CBT vs. applied relaxation, psychodynamic, biofeedback, or supportive therapy for GAD	Inconclusive
Normann et al. (2014)	Within	CBT vs. metacognitive therapy for anxiety and depression	No
Tolin (2014)	Within	CBT vs. others for anxiety disorders	No
Turner et al. (2014)	Within	CBT vs. befriending vs. cognitive remediation vs. psychoeducation vs. social skills training vs. supportive training for schizophrenia	No
Chen et al. (2015)	Within	CBT vs. EMDR for PTSD	No
Cristea et al. (2015)	Within	CBT vs. behavioural activation vs. supportive therapy vs. "other" therapies for cognitive change in depression	Yes
Lee et al. (2015)	Within	ACT vs. CBT for substance use	Yes
Tran & Gregor (2016)	Within	Trauma-focused therapies (e.g. TF-CBT, EMDR, prolonged exposure) vs. non-trauma-focused therapies (e.g. present-centred therapy) for PTSD	No
Linardon & Brennan (2017)	Within	CBT vs. other therapy on quality of life in eating disorders	No

Table 1

Meta-Analytic Conclusions from Direct Comparisons of Psychotherapies, November 1997 to June 2025

Study	Across or within disorders	Direct comparators	Authors' conclusions support treatment equivalency/DBV
Yulish et al. (2017)	Within	Psychotherapies (undifferentiated) for anxiety disorders	No
Grenon et al. (2018)	Within	Non-bona fide therapies vs. non-CBT therapies vs. CBT for eating disorders	Yes
Linardon (2018)	Within	CBT vs. behavioural weight loss vs. CBT+behavioural weight loss vs. IPT vs. behaviour therapy vs. DBT vs. mindfulness vs. psychodynamic therapy vs. supportive therapy for binge eating disorder	No
Singh & Gorey (2018)	Within	CBT vs. mindfulness for anxiety	Yes
Jones et al. (2018)	Within	CBT + standard care vs. other therapies + standard care for schizophrenia	Yes
Podina et al. (2019)	Within	Cognitive- vs. exposure-based CBT for anxiety disorders	No (for individual therapy)
Turner et al. (2020)	Within	CBT vs. other therapies for psychosis	No
Lewis et al. (2020)	Within	CBT with trauma focus vs. CBT without trauma focus vs. present centred therapy vs. supportive counselling vs. psychodynamic therapy vs. IPT vs. cognitive processing therapy vs. prolonged exposure vs. dialogical exposure therapy	No

Table 1

Meta-Analytic Conclusions from Direct Comparisons of Psychotherapies, November 1997 to June 2025

Study	Across or within disorders	Direct comparators	Authors' conclusions support treatment equivalency/DBV
		vs. EMDR vs. EFT vs. relaxation training vs. written exposure therapy vs. psychoeducation vs. virtual reality therapy for PTSD	
Barnicot et al. (2020)	Within	Psychotherapies (ACT, CBT, EMDR, IPT, metacognitive training, motivational, psychoeducation) vs. various nonpharmacological active interventions for psychosis	No for relapse, re-hospitalization, or treatment compliance; Yes for symptoms and functioning
Mavranezouli et al. (2020)	Within	TF-CBT vs. non-TF CBT vs. relaxation vs. EMDR vs. IPT vs. supportive counselling vs. present-centred therapy vs. combined somatic/cognitive therapies for PTSD	No
Ginley et al. (2021)	Within	Contingency management vs. community-based comprehensive therapy, non-specific therapy, or protocol-focused specific therapy (e.g., CBT)	No
Furakawa et al. (2021)	Within	Internet-delivered CBTs (third-wave cognitive behavioural therapy vs. behavioural activation vs. psychoeducation vs. problem-solving therapy) for depression	Yes (BA vs. Psychoeducation); No (all others)
Hoppen et al., (2022)	Within	Technology-delivered CBTs vs. other treatments (supportive therapy, relaxation, ERP, CBT, progressive muscle relaxation) for OCD	Yes
Fluckiger (2022)	Within	Bona fide therapies for GAD (Individual CBT vs. ACT vs. acceptance-based behavioral therapy vs. analytic psychotherapy vs.	Yes (primary measures); No (secondary measures; standard CBT vs.

Table 1

Meta-Analytic Conclusions from Direct Comparisons of Psychotherapies, November 1997 to June 2025

Study	Across or within disorders	Direct comparators	Authors' conclusions support treatment equivalency/DBV
		adherence priming implemented CBT vs. applied relaxation vs. behavior therapy vs. computer-assisted group CBT vs. group CBT vs. CBT plus interpersonal and emotional processing therapy vs. CBT plus interpersonal therapy vs. CBT plus supportive listening vs. CBT-well-being therapy vs. CT vs. DBT vs. intolerance-of-uncertainty therapy vs. meta-cognitive therapy vs. Motivational Interviewing-CBT vs. "Prolonged Focus on Change check-in phase implemented" CBT vs. "resource priming implemented" CBT vs. "state of the art check-in phase implemented" CBT vs. short-term psychodynamic psychotherapy vs. "supportive resource priming implemented" CBT vs. worry exposure)	augmented integrative CBT)
Papola et al. (2022)	Within	Psychotherapies for GAD (CBT vs. behaviour therapy vs. cognitive restructuring vs. psychodynamic therapy vs. psychoeducation vs. relaxation vs. supportive therapies vs. third-wave CBTs)	Yes
Jeong (2023)	Within	Psychotherapies for prevention of suicide re-attempt in psychiatric emergency (CBT vs. skills-based treatment vs. family-based therapy vs. supportive therapy)	Yes

Table 1

Meta-Analytic Conclusions from Direct Comparisons of Psychotherapies, November 1997 to June 2025

Study	Across or within disorders	Direct comparators	Authors' conclusions support treatment equivalency/DBV
Wen et al. (2023)	Within	ACT vs. 12-step facilitation vs. CBT vs. contingency management for opioid dependency (positive urine test rate)	Yes
Caselli et al. (2023)	Within	Short-term psychodynamic therapy vs. CBT vs. supportive therapy for depressive disorders	No
Papola et al. (2024)	Within	Behaviour therapy vs. CBT vs. cognitive restructuring vs. psychodynamic therapy vs. psychoeducation vs. relaxation vs. supportive therapy vs. third-wave CBTs for GAD	Yes
Smith & Hewitt (2024)	Within	CBT vs. psychodynamic therapy for depressive disorders	Yes (post-treatment); Inconclusive (follow-up)
Gupta (2024)	Within	EMDR vs TF-CBT for PTSD	No
Halicka (2024)	Within	MET-CBT vs. MET-CBT with affect management vs. MET-CBT + abstinence-based contingency management vs. METC-CBT + attendance-based contingency management for cannabis use disorder	Yes

Note: Conclusions are presented according to study authors' interpretations. ACT = acceptance and commitment therapy. CBT = cognitive behaviour therapy. CT = cognitive therapy. DBT = dialectical behaviour therapy. EMDR = eye-movement desensitization and reprocessing. GAD = generalized anxiety disorder. IPT = interpersonal therapy. OCD = obsessive-compulsive disorder. MET-CBT = motivation-enhanced CBT. PTSD = post-traumatic stress disorder. TF-CBT = trauma-focused CBT.

As Table 1 indicates, meta-analyses of direct comparisons in the past two and a half decades have, largely speaking, yielded conflicting results with respect to the DBV, with approximately one-third of the 53 identified meta-analytic studies supporting the DBV,

approximately 55% rejecting it, and approximately 10% having mixed or inconclusive findings. Of the identified studies, most were specific to one particular disorder or class of disorders. However, seven meta-analytic studies examined the relative efficacy of psychotherapies across disorders, providing a comprehensive evaluation of the DBV. Consequently, it is worth discussing them in some detail (here, they are presented in order of date of publication).[3]

Tolin (2010). Nearly fifteen years after Wampold et al.'s (1997) comprehensive meta-analysis exploring effect sizes without categorization into therapy classes, Tolin (2010) undertook a new comprehensive meta-analysis of the DBV, but using a very different approach.

Tolin focused on a single therapy class, CBT, and compared all other treatments to it for a variety of different psychological conditions, classifying therapies into CBT, psychodynamic therapies, interpersonal therapy, supportive therapies, and "other" therapies. Studies were identified using a database search. Studies were limited to randomized controlled trials with direct comparisons of *bona fide* psychotherapy, thereby accounting for "intent-to-fail" conditions (i.e., pseudo-treatments that would not be delivered in real-life psychotherapy). Analyses were undertaken both with post-treatment and follow-up data, with separate analyses for both symptom-specific and global treatment measures. A random-effects model was used to calculate the relative effects between CBT and other classes of therapy, and between-study heterogeneity was evaluated with significance tests; however, this heterogeneity was not quantified. The author considered whether studies had used completer analyses (i.e., removed the data of participants

[3] Recently, Pim Cuijpers's research group has released a series of meta-analyses of psychotherapy that cut across disorders (e.g., Cuijpers, Harrer, Miguel, Ciharova, Papola, et al., 2025; Harrer et al., 2025; Jiménez-Orenga et al., 2025). However, they have so far made use of indirect comparisons to a control group rather than direct comparisons. In particular, the paper by Cuijpers, Miguel, Ciharova, Harrer, Basic, et al., 2024 ("Absolute and relative outcomes of psychotherapies for eight mental disorders: A systematic review and meta-analysis") suggests by its title that it will examine relative efficacy. However, the title refers to meta-analyzing both (1) absolute measures of response and (2) relative *risk*—not to conducting a meta-analysis of relative efficacy. Accordingly, these studies were not reviewed in detail here.

who had dropped out and only used data from participants who completed the trial) or used intent-to-treat analyses (ITT; i.e., including the data from all participants so as to preserve randomization—such as by carrying the last observation forward to post-treatment), and used ITT analyses when both were reported. To ensure that multiple dependent effects (e.g., when the same set of participants complete two measures) were not erroneously treated as independent effects in the analysis, they were categorized separately where possible; where not possible (e.g., two measures of the same symptom from the same participants), an average effect was calculated. Publication bias was accounted for with Rosenthal's fail-safe N (Rosenthal, 1991). Meta-regression was used to test for the moderating effects of (1) the class of disorder in which the study was undertaken (anxiety disorders, depressive disorders, eating disorders, substance use disorders, psychotic disorders, developmental disorders, habit disorders, personality disorders, martial distress, academic problems, subclinical problems, and "other"); (2) researcher allegiance to a class of therapy; and (3) methodological quality.

Tolin (2010) reported significantly greater effects for CBT overall when compared to the other classes at post-therapy, $d = 0.22$, 95% CI [0.09, 0.35], six-month follow-up, $d = 0.47$, 95% CI [0.29, 0.66], and 1-year follow-up, $d = 0.34$, 95% CI [0.06, 0.62]. Subgroup analysis revealed significantly greater effects of CBT at post-therapy relative to psychodynamic therapies, $d = 0.28$, 95% CI [0.12, 0.44], and "other" therapies, $d = 0.22$, 95% CI [0.09, 0.35], but not relative to interpersonal or supportive therapies; analyses also indicated that the outcome with CBT was substantially greater than with psychodynamic therapies at six-month follow-up, $d = 0.50$, 95% CI [0.29, 0.71], and one-year follow-up, $d = 0.55$, 95% CI [0.30, 0.81]. When breaking down the post-therapy results by disorder, Tolin found that CBT outcomes were significantly superior to outcomes for all other therapy types for depression and anxiety disorders, but not for the other

conditions examined in the study. Moderator analyses found that the rating of the principal investigator's allegiance to a given therapy type significantly predicted study effect size, but that ratings of methodological quality did not. CBT was found to be superior on measures of primary symptoms, $d = 0.22$, 95% CI [0.09, 0.35], global symptoms, $d = 0.21$, 95% CI [0.01, 0.42], and general functioning, $d = 0.22$, 95% CI [0.09, 0.35], but not measures of comorbid symptoms, quality-of-life, self-concept, or social adjustment. Significant between-study heterogeneity was found for the overall post-treatment effect, $Q = 49.37$, and the overall 1-year follow-up effect, $Q = 16.46$, but not the 6-month follow-up, $Q = 21.98$; significant heterogeneity was also found for the effect with primary symptoms, $Q = 49.37$, but not the other measures. Most importantly, effects were not robust against the file-drawer effect, with the exception of the 6-month follow-up effects overall and relative to psychodynamic therapy.

The meta-analysis by Tolin (2010) included a number of methodological improvements relative to older meta-analyses—for example, comprehensively testing the DBV while making provisions to account for within-treatment class comparisons (i.e., by classifying treatments and including only between-class comparisons). Tolin also included tests of publication bias and tests of important potential moderators of heterogeneity (e.g., methodological quality, allegiance effects). However, this meta-analysis has been critiqued for its failure to include a number of key studies included in other relevant disorder-specific meta-analyses on the topic, particularly concerning anxiety disorders (Baardseth et al., 2013). The limited study pool is likely due to an inadequate search strategy: Tolin's database search relied on a small number of keywords used by the databases to code studies related to CBT, combined using the "AND" operator with terms used to tag studies related to non-CBT therapies; these were then further limited to those coded by the databases as RCTs or treatment outcome studies. These keywords do not capture relevant

studies that were not explicitly tagged by the databases and do not search for the key terms in important domains like the title or abstract. Additionally, there are questions about Tolin's classification scheme for what constitutes CBT, which included EMDR, despite its theoretical differences to CBT (although it has little difference in terms of techniques; e.g., Davidson & Parker, 2001).

Ruiz (2012). This meta-analysis evaluated head-to-head comparisons of acceptance and commitment therapy (ACT) relative to traditional CBT across disorders. Studies were identified using database searches, root-and-branch searches, reference list searches (i.e., of other meta-analyses), and a call for unpublished papers on ACT forums. Using a random-effects model, effects were aggregated for depression, anxiety, quality of life, and primary outcomes both at post-test and follow-up. Ruiz also examined the effect of changes in measures of (ostensibly) ACT-specific processes (e.g., cognitive defusion, experiential avoidance) and CBT-specific processes (e.g., dysfunctional attitudes or beliefs, anxiety sensitivity). Heterogeneity was assessed for significance and for the proportion of non-error heterogeneity, but not for the absolute amount of heterogeneity. The author calculated average effects to account for dependencies where they arose and used statistical techniques to account for publication biases. Only completer analyses were included. Moderators of between-study heterogeneity included anxiety/depression vs. other clinical problems, and CBT with cognitive interventions vs. without. Outliers, methodological quality and researcher allegiance were not accounted for.

The author found evidence that ACT was significantly more efficacious than CBT for primary outcomes across timepoints, $g = 0.40$, 95% CI [0.16, 0.64], at post-treatment, $g = 0.37$, 95% CI [0.12, 0.63], and at follow-up, $g = 0.42$, 95% CI [0.15, 0.69]. ACT had a significantly higher impact on ACT process measures than did CBT at combined timepoints, $g = 0.38$, 95% CI

[0.03, 0.73], and at post-treatment, $g = 0.45$, 95% CI [0.11, 0.79], but not at follow-up, $g = 0.01$, 95% CI [-0.13, 0.33]. The two did not differ significantly for depression outcomes, anxiety outcomes, or quality of life outcomes, or on CBT process measures. There was significant heterogeneity between studies for primary outcomes (all timepoints), depression outcomes (combined timepoints and post-treatment), anxiety outcomes (combined timepoints and post-treatment), quality of life outcomes (all timepoints), and ACT process measures (combined and post-treatment). Disorder class and presence of cognitive interventions in CBT did not moderate the effects. Publication bias was detected, and Duval and Tweedie's trim-and-fill procedure produced a corrected effect for primary measures of $g = 0.30$, 95% CI [0.06, 0.54].

A strength of this study is its distinguishing between therapy-specific process-based measures, a commendable decision that allows for determinations beyond relative efficacy (e.g., about differences in *how* therapies achieve their successes). It also provides a specific, clear, and straightforward test of relative psychotherapy efficacy by limiting its focus to head-to-head comparisons of just two psychotherapies (i.e., CBT and ACT). By the same token, this analysis is far from comprehensive—by design, it does not generalize to non-CBT/ACT psychotherapies, and only addresses to a limited degree the question of whether, overall, all psychotherapies are equally effective. Additionally, this study did not account for "intent-to-fail" treatments by limiting comparisons to *bona fide* psychotherapies. Indeed, some of the included comparisons appear to have been between ACT and behavioural *techniques* (e.g., progressive relaxation training, problem solving) rather than comprehensive CBT (and indeed, one comparison was to motivational interviewing, an evidence-based person-centred procedure that is not by itself CBT). By relying on completer analyses only, the author undertook a rather liberal test of differences. Furthermore, although the author interpreted the results as being robust against a

file-drawer effect, this does not appear to be the case; the author calculated a Rosenthal's fail-safe number of 78 for the primary measures effect, and by convention, a robust fail-safe number for the quantity of effects included in the meta-analysis would exceed 85 (Rosenthal, 1991). Lastly, the author did not account for methodological quality or researcher allegiance—potentially significant confounds, given that most included studies were undertaken by researchers interested in ACT.

Marcus et al. (2014). Marcus et al.'s (2014) meta-analysis updated the study by Wampold et al. (1997). Like Wampold et al. (1997), the authors randomly assigned directionality to the effect sizes of direct comparisons of *bona fide* psychotherapies—without reference to treatment classification or to clinical problem—and used a heterogeneity test to assess whether overall, effect sizes were heterogeneous around zero (which would falsify the DBV). Also like the previous study, they meta-analyzed the absolute differences between treatments to estimate the effect size of the overall difference. However, the authors modified this technique by analyzing disorder-specific measures separately from global measures and only included treatment studies that had participants drawn from clinical populations. Furthermore, unlike Wampold et al. (1997), Marcus et al. accounted for potential differences between theoretical orientations by undertaking pairwise comparisons of CBT to other different therapy categories (ACT, behavioural, IPT, psychodynamic, and other). They also tested the moderating role of disorder on the relative efficacy of CBT (anxiety, depression, eating disorder, and other disorder). To test for heterogeneity, the authors tested significance and for the proportion of non-error heterogeneity (but not absolute quantities of heterogeneity). They located their studies using a manual search of key journals. Appropriate random-effects models were used. The authors accounted for dependent effects by creating an average effect size, and undertook

statistical tests to account for publication bias, but did not account for ITT versus completer analyses, methodological quality, or allegiance effects. Outliers were accounted for statistically (e.g., analyses of whether removing any one study affected the results).

In contrast to the results found by Wampold et al. (1997), and inconsistent with the DBV, Marcus et al. (2014) found significant heterogeneity for primary symptom measures at both termination, $Q(49) = 107.48$, $p < .001$, $I^2 = 54.41\%$, and at follow-up, $Q(38) = 63.02$, $p = .007$, $I^2 = 39.70\%$. No significant between-treatment heterogeneity was found for secondary measures at either time point. The researchers found that, at termination, the average absolute effect size difference among treatments for primary measures was $d = 0.29$ (no confidence interval reported). They also noted that this difference at pre-treatment was $d = 0.09$ when it should have *ex hypothesi* been 0. They suggested, therefore, the treatment effect size may have been closer to $d = 0.20$ than to 0.29. Using Monte Carlo methods, they determined that this effect size difference was statistically significant at the $p < .001$ level. The authors also followed the strategy used by Tolin (2010) and Baardseth et al. (2013) in undertaking a contrast of CBT with other therapies, finding a superiority for CBT on primary measures at termination, $d = 0.16$, $p < .001$, 95% CI [0.07, 0.26], and at follow-up, $d = 0.16$, $p = .002$, 95% CI [0.06, 0.26]. Consistent with their earlier analyses, no CBT superiority was found at either time point on the secondary measures.

This study substantially improves upon the Wampold et al. (1997) methodology with appropriate tests to account for sources of heterogeneity (e.g., testing the moderating role of disorder and treatment category). The use of Monte Carlo methods to test the significance of the absolute value meta-analysis was novel and commendable. The authors' judicious use of outlier analyses would render their meta-analysis resilient to later criticisms of biased analysis (e.g.,

Wampold et al., 2017). However, given the role that researcher allegiance has been shown to play elsewhere in the literature, testing for the moderating role of allegiance would have strengthened the study, and the study is lacking an assessment of methodological quality (either as a moderator or through a risk-of-bias assessment). The study's reliance on a limited pool of journals is practical, but it is a potential source of systematic bias (e.g., Leichsenring et al., 2017).

Kivlighan et al. (2015). The authors of this study aimed to test the superiority of psychodynamic therapy in the long-term for a number of outcome domains (symptom-specific, global functioning, and personality outcomes, as well as a combination of all outcomes). The authors used database searches, reference list searches, and consultation with experts to identify RCTs comparing psychodynamic to non-psychodynamic therapies. To account for "intent-to-fail" control treatments, they limited the studies to those meeting the criteria for *bona fide* psychotherapy (Wampold et al., 1997). Treatments were classified as psychodynamic therapy or not by consultation with psychodynamic experts; other therapies were not classified in any other way. Studies were also not classified by their clinical population (i.e., disorder was not accounted for). Multiple dependent effects were combined into an average effect to maintain the assumption of independence. Kivlighan et al. (2015) used multilevel longitudinal meta-analysis to account for multiple treatment assessment times within studies; by using hierarchical linear modelling (HLM), the authors were able to test the growth of the effect of psychodynamic therapy over time vis-à-vis other therapies while controlling for number of sessions, as well as test for treatment differences at post-test. The authors used a random effects model and tested the significance and the proportion of non-random heterogeneity (but did not evaluate the absolute

quantity). The authors also tested the moderating role of researcher allegiance. Publication bias and methodological quality were not accounted for.

The authors found no significant differences between psychodynamic therapies and non-psychodynamic therapies at post-test, for any measure. Effects were all less than $g = 0.10$ and were generally less than $g = 0.01$. There were also no significant differences between psychodynamic and non-psychodynamic therapies in terms of the change in their treatment effects over time for any measure, and there was no significant heterogeneity between studies for change over time any type of measure. Allegiance significantly moderated change in disorder-specific outcomes, such that stronger psychodynamic allegiance led to *decreased* growth in psychodynamic effects. However, given (a) that no group differences and no heterogeneity between studies were found for the change in disorder-specific measures and (b) the highly unusual relation suggested by this finding, there is reason to doubt the robustness or accuracy of this finding. At post-test, there was significant heterogeneity among the effects for non-targeted outcomes and personality outcomes (but not for disorder-specific or combined outcomes); allegiance did not significantly moderate any of the effects.

The unexplained heterogeneity underscores a major limitation of this study—namely, it did little to reduce or account for obvious potential sources of heterogeneity among the study effects, such as the theoretical orientation of the comparator treatment, disorder, or methodological quality. The authors emphasized that most of the comparisons included in the analysis were with CBT, going so far as to say that "the results of the present meta-analysis are more generalizable to the comparison of psychodynamic treatments and bona fide cognitive-behavioral treatments, rather than to any bona fide alternative treatment that may be only minimally represented in the current analysis" (Kivlighan et al., 2015, p. 11). If the authors

believed this, it would have been valuable to analyze CBT versus psychodynamic therapies separately, so that such determinations could actually be made. This is especially important given that some of the largest included studies in terms of sample size compared psychodynamic therapy to a non-CBT therapy; because summary effect sizes are weighted based on sample size, these would have had an outsized impact on the meta-analytic results. For example, one included dataset (n = 326) involved comparisons of an average of 9.8 sessions of solution-focused therapy to an average of *232* sessions of long-term psychodynamic psychotherapy. Although measures were taken to statistically control for number of sessions, these regression analyses may have been insufficient to account for such an outlier (as most other sample sizes were much more evenly weighted between groups). As no steps were taken to assess or account for outliers—another major limitation of this study—it is not obvious what to ultimately make of the results. A major methodological strength, however, was this study's consultation with psychodynamic experts for their classification, helping to bolster the validity of their classification system.

Goldberg et al. (2018). The authors of this study undertook a meta-analysis of RCTs comparing mindfulness interventions to other types of therapy across psychiatric disorders. They accounted for comparisons to intent-to-fail conditions by including the type of control condition as a moderator—levels included "no treatment," "minimal treatment," "nonspecific active controls" (i.e., intent-to-fail) and "specific active controls" (i.e., *bona fide* psychotherapy). They also separately examined evidence-based treatments (EBTs; i.e., as recognized by American Psychological Association [APA] Division 12 or a similarly relevant organization for a particular clinical issue; e.g., CBT for insomnia by APA Division 12 or smoking cessation treatment by the American Lung Association). Studies were identified with database and reference list searches. The authors undertook comparisons on symptom-specific measures only, but they included both

post-treatment and follow-up comparisons. Dependent effects (e.g., multiple measures from the same sample) were treated as independent effects, potentially impacting the results; however, these effects were examined with sensitivity analyses, and did not appear to strongly impact the conclusions. The authors used a random-effects meta-analytic model to produce a weighted summary effect size. They tested for statistically significant heterogeneity between studies, as well as the proportion of heterogeneity not due to sampling error; however, they did not examine the absolute quantity of heterogeneity. Using meta-regression, they examined the degree to which between-study heterogeneity could be explained by moderators such as intent-to-treat vs. completer analyses or the disorder examined in each study; however, they did not statistically examine the impact of researcher allegiance. Methodological quality was examined using a risk-of-bias analysis, and statistical procedures were used to examine publication bias.

Goldberg et al.'s meta-analysis found that mindfulness interventions were more efficacious than non-specific active controls at termination (d = 0.35) and follow-up (d = 0.52); they were also more efficacious than *bona fide* psychotherapies at termination (d = 0.23) and follow-up (d = 0.29). They did not differ from EBTs (d = -0.004 at termination, d = 0.09 at follow-up). Moderator analysis revealed that mindfulness was superior for smoking relative to EBTs but not for anxiety or depression; disorder was not a significant moderator for other *bona fide* therapies. These outcomes were generally unchanged by removing dependent effects and by accounting for publication bias; ITT vs. completer analyses did not moderate outcome effects.

Despite the generally excellent design of this meta-analysis, one possible concern is the studies that were included in the analysis. Although the study was billed as a test of the efficacy of mindfulness for various psychiatric disorders, the authors included studies for problems such as food cravings, stress eating, headaches, fibromyalgia and pain, eating disorder prevention, and

prevention of depression relapse (i.e., not active eating disorders or depression). Some of the outcomes included were of questionable relevance, including body awareness (rather than pain reduction) for chronic pain (see Goldberg et al., 2018, Supplemental Materials Table 4a). The study also did not address the relative effects of mindfulness versus other therapies for non-disorder specific measures. Lastly, this study did not provide information about the relative effects of different theoretical orientations; however, by classifying comparator treatments by their degree of purported strength (i.e., EBTs, specific active controls, non-specific active controls) the authors nonetheless provided a meaningful way to parse the heterogeneity of the "other therapy" class.

Munder et al. (2019). Munder et al. (2019) compared standard versions of existing therapies, such as CBT and psychodynamic therapy, to adaptations of those same therapies that focus on enhancing patients' pre-existing strengths (as opposed to the standard approach of focusing on tackling patients' dysfunctions and deficits). The authors located their studies using a systematic search of Cochrane, PsycINFO, and PSYNDEX, as well as by using a stem-and-branch search of relevant studies and by contacting experts. One subgroup in their meta-analysis, which this review will focus on, looked at head-to-head comparisons of these alternative therapies from randomized controlled trials; other groups in their analyses included comparisons to waitlist and those from non-randomized trials. The authors excluded healthy populations (e.g., students) from their analysis, limiting the included studies to those with participants with diagnosed mental health disorders or "clearly distressed populations" (e.g., trauma-exposed). The authors reported a liberal approach to the setting and modality of therapy (i.e., group vs. individual), but they required that the therapy have "some sort" of individualization to the patient and "some sort" of therapist contact; no further definition of either was provided. They explored

the moderating role of outcome measure, timepoint (post-therapy or follow-up), therapy family, methodological quality (operationalized using the Cochrane RoB2 measure) and researcher allegiance (operationalized as whether the author developed one of the therapies in question). The authors used a random-effects model and calculated two effect size measures: the standard Hedges *g*, and an adaptation that accounts for pre-treatment differences ($g_{\text{pre-post with control}}$, g_{ppwc}). The authors also appraised risk of publication bias using a funnel plot and Egger's test.

The meta-analysis ultimately included five comparisons of active comparators drawn from RCTs, and found that across all outcomes at post-treatment, there was a difference of g = 0.25 and g_{ppwc} = 0.47 in favour of strengths-based adaptations of standard therapies relative to the standard therapies. The authors found that, according to Cochran's Q test, there was significant heterogeneity in the g_{ppwc} effect size, but not in the standard g effect size. The authors used the I^2 metric to describe the degree of heterogeneity; however, I^2 is a relative measure, and the authors did not provide an absolute measure to quantify the extent of the dispersion of the effects. The authors reported an effect size for (1) the follow-up timepoint and (2) various outcome measures (i.e., primary symptoms vs. interpersonal functioning vs. quality-of-life outcomes). However, there were fewer than three comparisons used to derive these effects, and they all included non-randomized trials. No heterogeneity tests were reported for the follow-up effect. The authors reported an asymmetric funnel plot, but no significant publication bias according to Egger's test.

This study provided an interesting test of the DBV in a manner not frequently seen in the meta-analytic literature: namely, a meta-analytic test of differences between treatments that is orthogonal to traditional classifications according to therapy families. Such a test is in line with recent trends in the theoretical literature towards focusing on cross-cutting therapy processes

rather than schools of therapy (e.g., Hofmann & Hayes, 2018). However, the authors reported that they coded effects by the traditional therapy families to which each intervention belonged, but reported no data associated with this coding (e.g., sensitivity or moderator analyses); such an omission raises questions about the potential for a reporting bias. Given the likely small number of primary studies focusing on strengths-based adaptations, it is unsurprising that the meta-analysis contained five analyses. Moreover, given the overall low methodological quality of psychotherapy RCTs (e.g., Cuijpers, Harrer, Miguel, Ciharova, & Karyotaki, 2025), it is unsurprising that the authors classified all included studies as either "high" or "unclear" risk of bias. However, both the small number of studies and their high risk of bias suggests that the conclusions must be interpreted with caution. Additionally, the authors' definition of researcher allegiance—as "present" if the authors developed the intervention—is narrow relative to other definitions (e.g., Miller et al., 2018; Yulish et al., 2017); this presents the risk that the authors underestimated the impact of researcher allegiance. Even so, the authors indicated that all included studies had either "present" or "unclear" researcher allegiance (note that the authors did not directly state what qualified as "unclear" allegiance)—suggesting a further important caveat when interpreting the results. Lastly, the somewhat nebulous notion of "strengths-based interventions" makes it plausible that the authors' search terms were inadequate to capture a representative selection of relevant primary studies, especially when the search was not limited to specific disorders. The authors' choice to supplement their search algorithm by using stem-and-branch and expert consultation strategies is a methodological strength that likely helped to compensate for this issue.

Linardon et al. (2019). Linardon et al. (2019) examined the effect of IPT in head-to-head comparisons with other psychological interventions across mental health disorders. They

used dropout from therapy as their measure of effect size (i.e., lower percentage of dropout indicating a more efficacious therapy). Primary studies were identified using a keyword/title/abstract search of the CINAHL, EMBASE, Medline, and PsycINFO databases, supplemented with root-and-branch searches of other trials and reviews. The authors used a random effects model and incorporated analyses of publication bias and methodological quality (RoB2). They also analyzed the impact of the following subgroups on effect size: various definitions of "dropout"; treatment format/modality; comparator treatment; sample age; session length; and whether comorbid substance use or personality disorders were excluded from the sample. In addition to the cross-disorder meta-analysis, the authors also undertook separate meta-analyses for various disorders individually. The cross-disorder analysis revealed significantly greater dropout in non-IPT therapies (approximately 25% vs. 18%); this pattern held for IPT vs. CBT and IPT vs. non-specific supportive therapies. Moreover, a similar pattern held when analyzing dropout in the anxiety disorder meta-analysis. For depressive disorders, the pattern was also similar, but the specific comparison of IPT vs. CBT was no longer significant. For eating disorders, there were no significant differences detected between IPT and other therapies. For the cross-disorder analysis, the authors undertook a sensitivity analysis to compare the results from (a) studies providing a direct statement that participants dropped from the treatment (as opposed to from the study more generally) to (b) studies using any definition of dropout. The main results were largely robust when comparing these definitions of dropout.

This study's approach provides an interesting perspective on the DBV: rather than comparing treatments on their benefits (e.g., symptom reduction, increases in well-being), their focus on dropout as an outcome explores the degree to which treatments differ in acceptability, and hence the potentially undesirable characteristics that presumably drive dropout. The negative

effects of therapy are an increasingly recognized as a playing an important role in patients' experiences of psychotherapy (e.g., Klatte et al., 2023; McQuaid et al., 2021); this study is consonant with that literature. Moreover, dropout has the additional helpful characteristic of being an objective measure of therapy acceptability. However, it is also difficult to appraise the results of Linardon et al. (2019) due to important information that was not reported. For example, the authors indicated that they collected data on methodological bias, but no data were not reported to indicate the impact of these ratings on the effect sizes. Similarly, the authors did not report *Q* tests for the heterogeneity of their head-to-head comparisons; although they reported the measure of relative heterogeneity, I^2, they did not report the relevant metric for the absolute amount of heterogeneity, τ. At least one of these measures would be needed to determine the presence or extent of dispersion in the data, and hence how to interpret the effect sizes. Additionally, although the authors completed publication bias analyses, the impact of publication bias on any of the head-to-head comparisons was not reported. In addition to these omissions, many of these studies were undertaken by researchers who were associated with the development of IPT; nevertheless, no analysis of the impact of researcher allegiance was included. Finally, some of the subgroup analyses were underpowered by the standard rules of thumb (i.e., less than 10 studies per group; e.g., the IPT vs. CBT analysis for eating disorders).

Unaddressed Issues: The Dodo Bird at 45

Four-and-a-half decades after Smith et al. (1980), the meta-analytic evidence on psychotherapy efficacy is mixed. Consensus appears to have been reached for certain key issues—for example, little question remains as to whether *bona fide* psychotherapeutic treatments are effective or ineffective, as across studies, all such treatments seem to provide benefit over and above control conditions (although the size of this effect varies on the basis of a

study's risk of bias; Cuijpers et al., 2019). However, disagreement still remains about the efficacy of psychotherapies relative to one another. It is the aim of this dissertation to address (1) why these disagreements occur, and in light of these issues, (2) undertake new meta-analytic evaluations that address these sources of disagreement.

The first paper in this dissertation is a methodologically focused review, aimed at describing in detail the heterogeneity of methods found in the meta-analytic literature on psychotherapy equivalence. If significant variation is seen in the methodology of these studies, then it stands to reason that this may be contributing to the diversity among the studies' conclusions. For example, certain meta-analyses may be comparing the effects of therapies relative to waitlist, whereas others may be examining direct psychotherapy comparisons; some meta-analyses might be primarily assessing overall heterogeneity in the literature, whereas others might be primarily comparing effect sizes; or, some might make more comprehensive efforts to account for bias (e.g., publication, allegiance) than others. The methodological review would be aimed at evaluating the impact of these factors in the literature and commenting on the appropriateness of certain methodological decisions relative to others in terms of assessing psychotherapeutic equivalence.

The second, meta-analytic paper addresses some of the methodological issues identified in the review. In this study, the DBV is approached using different methodological tactics, which allows for determinations as to whether these various strategies impact the meta-analytic "verdict." First, the DBV is appraised by (1) replicating the methods of Wampold et al. (1997)—that is, (a) with a meta-analysis of randomly-signed effect sizes, using heterogeneity measures as a global test of the overall differences between treatments, and (b) with a meta-analysis of the absolute effect sizes between treatments, to test the magnitude of the difference. In line with

Marcus et al. (2014), this study then undertakes (2) a standard meta-analysis, selecting one treatment (e.g., CBT) as a comparator class and using this to calculate standard effect sizes. Meta-regression techniques are used to assess how the summary effect size is impacted by variables such as the clinical population, the studies' risk of methodological bias, and whether the studies' comparators come from different therapy classes (e.g., CBT versus psychodynamic) or from the same one (e.g., CBT versus another CBT).

The analytic approach in this dissertation—to meta-analytically synthesize numerous primary studies of psychotherapy efficacy, with an eye to evaluating whether the literature supports the DBV—is not novel, as the historical review demonstrates. Indeed, it is in one sense quite a traditional approach (e.g., Smith & Glass, 1977). Its novelty does not lie in innovation, but in combination: in the methodological review, meritorious methodological strategies are identified, and in the meta-analytic paper, several such strategies are combined that have heretofore been found only in some meta-analyses, but never all together.

Where Wonderland Ends: The Bounds of a DBV Meta-Analysis

Many long-standing controversies in psychology, such as the one surrounding the DBV, can often be traced to disputes not about empirical findings *per se*, but about how those findings are framed, interpreted, or operationally defined. It may therefore be prudent to delineate the scope of the DBV controversy as it was be addressed in this dissertation. The central question addressed by this dissertation, framed in two parts, is this: (1) Across head-to-head studies of psychotherapy outcome, are all approaches to psychotherapy equally efficacious, or are some approaches more efficacious than others? (2) If so, for what clinical conditions are some approaches to psychotherapy more efficacious than others?

Empirical versus Evaluative Inquiry. The first issue to notice about this question is that it is empirical. It is concerned with the relative efficacy among approaches to psychotherapy, wherein relative efficacy is defined in terms of some relevant measured outcome (e.g., a reduction on a measure of symptom distress or an increase in a measure of general well-being). It is not concerned with adjudicating which approach to therapy is "better" or "worse" than others; indeed, one might imagine that there could be a relatively less efficacious approach to therapy which might have several other features that recommend its use (e.g., low-cost, highly efficient, easy to disseminate and implement). Although these other factors remain important, this dissertation was focused exclusively on efficacy—that is, controlling for as many extraneous factors as possible, does one approach to therapy lead to better mental health outcomes for patients than others?

Relative versus Absolute Efficacy. The second issue to note is that this question is focused on *relative* efficacy, not absolute efficacy. As has been discussed (see "History," above), questions of relative and absolute efficacy have historically been conflated by arguments that (1) drew a distinction between behaviour therapies and other psychotherapies and (2) contended non-behavioural therapies provided no more benefit than the rate of spontaneous remission (e.g., Eysenck, 1952). This historical usage of the term "psychotherapy" is, however, not standard today, and this question would today be understood as one concerning relative efficacy. It is important to maintain the distinctiveness of questions that concern relative and absolute efficacy, as it is entirely possible for all psychotherapies to be effective without their being *equally* effective. Indeed, there is consistent evidence that numerous different approaches to psychotherapy can provide efficacious treatment in an absolute sense for adults' clinical

psychological difficulties (e.g., Butler et al., 2006; Hunsley et al., 2014; Lambert, 2013; Leichsenring et al., 2004).

Orthogonality to the "Common Factors" Argument. A third issue to notice is that this question makes no direct reference to the relative importance of common or specific factors in therapy. This point bears emphasizing as, since its inception, the DBV debate and the debate about common and specific factors in therapy have been closely linked (see "History," above). The phrase "common factors" generally denotes processes of therapeutic change that are transtheoretical and which transcend the specific explanatory frameworks of therapy families—but the phrase can have two different connotations. It may be used, on the one hand, to mean therapeutic change mechanisms that are not related to specific therapeutic procedures but rather to more interpersonal factors (e.g., empathy; therapeutic alliance; persuasion and expectation). This is, roughly, the sort of sense in which the phrase has been used by Wampold et al. (2015), Lambert (2013) and Frank (1961; 1971). Alternatively, it may connote *any* transtheoretical change process, including those associated with specific change theories and therapeutic procedures (e.g., "experiencing avoided emotions in a safe context" as one of the change processes underlying techniques such as systematic exposure, mindfulness, and two-chair exercises). This sense of the term has been employed by Goldfried (1980), Norcross (2005), Prochaska and DiClemente (1979), and Hofmann and Hayes (2018). To clarify this ambiguity, "common factors" will here be used in Wampold's sense of "change principles unrelated to technique," whereas cross-cutting principles of change specifically related to technique will be referred to as "therapeutic processes." The general term for the category containing both "common factors" and "therapeutic processes" will be referred to as "transtheoretical factors."

In the latter-day literature on psychotherapy, proponents of the common factors explanation for the DBV generally hold that relative to common factors, the specific factors have little to do with psychotherapy efficacy (e.g., Wampold & Imel, 2015). According to this line of thinking, psychotherapies are either equally effective or not. If they are not equally effective then, according to this argument, this supports the position that certain therapeutic techniques or processes ("specific factors") are superior to others. However, if they are equally effective then, according to this argument, specific factors have little to do with the efficacy of therapy and, instead, suggests that "common factors" such as the therapeutic alliance and client expectation are primarily responsible for treatment efficacy (e.g., Wampold & Imel, 2015).

There is certainly a strong empirical case to be made for the effect of various common factors on psychotherapeutic outcomes (e.g., Horvath, , 2011; Graves et al., 2017); however, the view that the DBV bears directly on the argument for the primacy of "common factors" over "specific factors" is far from straightforward. The previously sketched argument contrasts two competing options: either (1) specific psychological theories propose theory-specific change processes which dictate theory specific techniques, leading to therapeutic benefit, or (2) a common factors theory proposes non-theory-specific common change processes, which manifest as non-technical and interpersonal therapeutic activities, leading to therapeutic benefit (e.g., Frank, 1961, 1971; Wampold et al., 2015). However, this dichotomy does not exhaust the space of options. One might consider a more nuanced conception of specific psychological theories than what was outlined in (1) above, as most cognitive-behavioural, psychodynamic, humanistic, or interpersonal approaches do not propose such a linear relation between theory, mechanism, and technique. For example, (3): even though specific psychological theories may argue that that change occurs primarily via their theory-specific mechanisms, they also generically acknowledge

the role of more general transtheoretical psychological processes of change of which their specific mechanism may simply be a special case, as well as the important role of common interpersonal and relational factors (e.g., Young & Beck, 1980). Indeed, certain change factors may be taken to be active but non-specific in one theoretical framework (e.g., empathy in some behavioural models) but active and theory-specific in others (e.g., empathy in person-centred models), rendering suspect the very division between common and specific factors. Additionally, one may consider yet further options, such as (4): an integrative theoretical framework that proposes both that (a) transtheoretical processes of psychological change dictate specific therapeutic actions and techniques, and (b) interpersonal common factors manifest as non-technical and relational therapeutic activities, all of which produce therapeutic benefit. Accordingly, even if the DBV were found to be strongly supported, and even if we were to take this to be disconfirmatory of (1), it would not strongly distinguish between (2), (3), or (4). Similarly, if the DBV were ultimately found to be contradicted by evidence and we were to take this as disconfirming (2), it would not strongly distinguish between (1), (3), and (4).

However, even considering only options (1) and (2), several alternative explanations could still exist for the DBV apart from the primacy of "common factors." For example, there could be "many roads to Rome": the therapies could be using different theory-specific mechanisms (or different "mixes" of theory-specific and shared, theory-incidental mechanisms) in a manner that leads to comparable outcomes—perhaps with each specific mechanism targeting a different, equally important, aspect of the psychological difficulty. Similarly, even if psychotherapies are not equally efficacious, a number of alternative explanations exist apart from the primacy of "specific factors." For example, all therapy approaches may be employing roughly the same change mechanisms, but they might employ better or worse models of clients'

problems (and therefore might be better or worse at correctly targeting their difficulties), leading to differential outcomes. Alternatively, it could be that "common factors" are the *only* mechanism of psychotherapeutic change, but certain therapies might be better than others at implementing them.

In short, because many possible explanations exist for the results of a DBV meta-analysis, any conclusions drawn about "common factors" would be under-determined by the meta-analytic data. That is not to argue that the common factors debate has *no* role in interpreting meta-analyses of relative psychotherapy efficacy; indeed, in the meta-analytic study, the role of common versus specific factors is explicitly contemplated as a speculative explanation for the results. Rather, the point is that direct process-outcome research is required to explain the mechanism producing equivalent or non-equivalent meta-analytic results, not the meta-analysis itself. Although the DBV may be historically closely related to the common factors debate, the two are not tied logically or conceptually—and a verdict on the former issue does not necessarily imply a verdict on the latter.

Psychotherapy Techniques, Theories, and Treatment Packages. When considering meta-analyses of the DBV, the majority of the relevant RCTs comparing one treatment to another will, in practice, be comparing delineated treatment *packages*: a particular combination of specific *techniques* (some more theory-specific, such as cognitive restructuring with thought records or two-chair exercises; some more non-specific, such as empathizing), along with specific *methods of delivery*, often endowed with particular psychological *processes* as their theoretical rationale, and labelled as representing a particular *theoretical framework* (e.g., cognitive-behavioural, psychodynamic). No one treatment package can be said to represent the totality of an overarching theoretical framework—a package involving systematic desensitization

and a package involving problem-solving techniques, for example, would be heterogeneous in their typical techniques and rationales, and neither could be said to fully represent the cognitive-behavioural approach. However, in aggregate, these packages may form a *therapy family* (e.g., cognitive behavioural therapies, psychodynamic therapies) of diverse, yet theoretically interrelated, sets of packaged techniques, delivery, and rationales. Such therapy families may be considered at an intermediate level of abstraction, generalizing away from individual techniques and the packages they comprise, but nevertheless more concrete than the purely abstract overarching theoretical frameworks (e.g., Goldfried, 1980).

In line with similar meta-analyses on this topic (e.g., Marcus et al., 2014), the meta-analytic study in this dissertation also aggregated RCTs that compare treatment *packages* (rather than, for example, directly comparing techniques). Accordingly, its conclusions about the DBV—the claim that all psychotherapies have equivalent outcomes—are most appropriately limited to a narrow sense of psychotherapy consisting of treatment packages and the therapy families they comprise. Its conclusions about these families cannot logically be extended to the specific techniques included in the packages, as the effects may be attributable to how the techniques are combined, implemented, or guided by theory. Similarly, its conclusions cannot be extended to the theoretical framework that unites the therapy family: the effects speak to the efficacy of how theoretical principles have been implemented in the packages, and not necessarily to the soundness of the principles themselves.

Problem Areas, Disorders, And the Nature of Psychotherapy. Answering this dissertation's central question—whether all approaches to psychotherapy are equally efficacious—also provides an opportunity to provide insight into the circumstances under which some might be more efficacious than others. The meta-analytic study assesses this with respect

to several moderating variables, including therapy family (specifically, whether effect sizes between therapy families differ from those within families), risk of bias, and researcher allegiance. It also appraises the effect of studies' populations in terms of psychological problems; in doing so, a particular perspective on the DBV is adopted, and accordingly its conclusions may be limited to that specific perspective (e.g., it may not apply to "verdicts" about psychotherapy for non-clinical problems).

There are several reasons for evaluating the impact of studies' populations. First, on methodological grounds: the aim of the dissertation, as previously stated, is to determine the relative efficacy of treatment packages associated with different psychotherapeutic families. Most treatment packages have been tested for the treatment of particular psychiatric diagnoses; indeed, many treatment packages were designed for particular diagnoses, incorporating specific procedures targeted towards particular purported psychopathological mechanisms (Barlow, 2004). Given the proliferation of treatments specific to particular clinical populations, testing for the moderating role of these differing populations on the effect size may be critical. Second, on theoretical grounds: although the DBV has historically been framed as an "overall" or "general" verdict on psychotherapy (e.g., Luborsky et al., 1975), there is much conceptual clarity offered by a further problem-specific breakdown. Indeed, it is arguably unclear exactly what it means for two psychotherapies to be equally effective or differentially effective *overall* without consideration of the problems, populations, and circumstances for which that therapy might be more or less efficacious (e.g., Paul, 1967). To paraphrase an example from Leykin and DeRubeis (2009), it would make little sense to ask "overall, are all drugs equally efficacious?" without asking "efficacious for what?" Extending the analogy: corticosteroids and antipsychotics would likely show no difference in efficacy overall (i.e., neither would likely show superiority for

ameliorating symptoms of diabetes, and both are approximately equally efficacious for the problems for which they are each prescribed; Leucht et al., 2015). However, one certainly expects corticosteroids to outperform antipsychotics at reducing symptoms of inflammation, and the antipsychotics to be better than steroids for schizophrenia symptoms. The same type of distinction may be true of psychotherapies, thus making a problem-based breakdown conceptually important.

What constitutes a relevant problem area has not been consistently agreed upon in the DBV meta-analytic literature. Some have confined their analysis to diagnosed disorders, whereas others have included in their tests of psychotherapy equivalence such problems such as high school underachievement and decisional conflict (e.g., Wampold et al., 1997). This meta-analytic study took the former approach, analyzing the impact of superordinate disorder categories (e.g., depressive disorders, anxiety disorders). This decision was driven, in part, by the aforementioned methodological reason: *pace* important exceptions (e.g., Riley et al., 2007), the majority of RCTs on psychotherapy that identify a clinical population do so in terms of psychological diagnoses (e.g., in terms of Diagnostic and Statistical Manual of Mental Disorders [DSM]- or International Classification of Diseases [ICD]-defined categories). However, this decision was also partly driven by the view that the DBV is primarily of relevance for questions of public health (specifically, the question as to whether certain treatment approaches can provide better outcomes than others for individuals suffering from diagnosable mental health problems). Given this perspective on the DBV, comparisons between treatments for nonclinical or subclinical psychological problems would arguably have limited relevance. Instead, this meta-analysis extended the framework of testing *bona fide* psychotherapies to the notion of testing them with respect to the treatment of *bona fide* clinical disorders. Because the DBV is a question of relative

psychotherapy efficacy, this stance may seem to be advocating for a particular perspective on psychotherapy—namely, that there is a distinction to be drawn between psychotherapies tailored to disorders and those that are not, and that the former are more relevant than the latter (e.g., Barlow, 2004; cf. Norcross, 1990). However, the perspective I intend to take in this dissertation is about the boundaries of the DBV rather than about psychotherapy itself (i.e., that the DBV is most practically relevant when concerned with mental health disorders). As a consequence of adopting this particular perspective, the conclusions reached by this dissertation may not necessarily generalize to other perspectives on the appropriate bounds of the DBV.

Putting It All Together

In summary: The central, two-part question to be answered in this dissertation is about the accuracy of the Dodo Bird Verdict. (1) Across head-to-head comparisons of *bona fide* approaches to psychotherapy, are all approaches equally efficacious, or are some approaches more efficacious than others? (2) If there are differences in efficacy, in what situations are some approaches to psychotherapy more efficacious than others? These questions are concerned with the relative efficacy of different psychotherapies (as opposed to absolute efficacy) and do not speak to the superiority or inferiority of different psychotherapies on any other dimension (including effectiveness in real world implementation). These questions, as construed in this dissertation, do not bear directly on the relative importance of "common" versus "specific" factors in therapy—instead, they address whether different families of treatment packages have shown efficacy differences, and are agnostic as to whether the theoretical mechanisms are "common" or "specific."

To answer these questions, methodological issues in DBV analyses were reviewed, and a meta-analytic study of the relative efficacy of psychotherapies was undertaken, examining

clinical trials that compare two psychotherapies for the treatment of adult mental disorders. Although these studies borrowed important methodological standards from older meta-analyses, such as limiting the inclusion of primary studies to those that involve direct comparisons of *bona fide* psychotherapies (Wampold et al., 1997), they also made use of meta-analytic methods that reflect advances in the state of the science in the past decades.

The New Caucus-Race:

Methodological Considerations for Meta-Analyses of Psychotherapy Outcome

Shawn G. Sanders and John Hunsley

University of Ottawa

Reference

Sanders, S. G., & Hunsley, J. (2018). The new Caucus-race: Methodological considerations for meta-analyses of psychotherapy outcome. *Canadian Psychology, 59*, 387–398. https://doi.org/10.1037/cap0000164

The New Caucus-Race:

Methodological Considerations for Meta-Analyses of Psychotherapy Outcome

In *Alice's Adventures in Wonderland*, the Dodo bird adjudicates a "Caucus-race:" a contest in which contestants begin at different starting points, run in different directions, and end whenever they please. The Dodo concludes the race and declares: "*Everybody* has won, and all must have prizes" (Carroll, 1865/2008). The Dodo bird's verdict (DBV) has been used to describe a particular conclusion about psychotherapy outcomes—namely, that all types of psychotherapy are equivalently effective (i.e., that all therapies have "won" and all of them deserve "prizes;" Luborsky, Singer, & Luborsky, 1975; Rosenzweig, 1936). Ironically, the DBV has acted as a starting gun for a new Caucus-race, with researchers going in different directions to find evidence in support of, or against, the verdict.

The primary strategy used by the runners of the new Caucus-race has been meta-analysis, with effect sizes from a number of relevant primary treatment studies collected, weighted by their precision, and then aggregated to produce a weighted average effect size. This strategy allows for a much more precise estimate of the overall effect than could be obtained by any one primary study, due to the increase in sample size and statistical power that arises from aggregation. As most primary studies in the psychological literature have low power to detect significant effects, this also yields a clearer overview of the literature than could be obtained by counting the number of studies with and without significant results (e.g., Borenstein, 2000).

Meta-analyses, however, are limited by both the nature of the primary studies included in the analyses and the manner in which the analyses are conducted. A high-quality meta-analysis should have a cogent justification for the studies included and excluded, and appropriate analyses should be used to combine data from included studies, evaluate possible biases in the data set,

and evaluate variables that may moderate the overall findings. However, meta-analyses on the DBV have historically been run as races with different starting points (i.e., variability in which studies are included), in different directions (i.e., variability in analytic strategies), and with arbitrary finish lines (i.e., variability in how the resulting mean effect sizes are interpreted)—much like the Dodo's Caucus-race. This has resulted in competing claims that all psychotherapies are essentially equivalent (e.g., Wampold & Imel, 2015), that certain therapies provide superior outcomes for some clinical conditions (e.g., Marcus, O'Connell, Norris, & Sawaqdeh, 2014), and that there may be insufficient evidence to decide one way or the other (e.g., Hunot, Churchill, Teixeira, & Silva de Lima, 2007).

In this article, we review critical methodological issues that have arisen in the context of using meta-analysis to evaluate the DBV. These issues, we contend, have interfered with researchers' ability to draw clear conclusions from the results of their statistical analyses and continue to impede attempts of mental health professionals to use the results of these meta-analytic studies to guide their clinical training and clinical service activities. The range of issues we address include questions of (a) how to aggregate effect sizes, (b) how to determine which outcomes from primary studies should be included in the analyses, (c) whether to conduct direct or indirect comparisons between treatments, (d) how best to conduct moderator analyses, and (e) how to provide contextually appropriate interpretations of effect sizes obtained in the meta-analysis. We have two goals in reviewing these issues. First, we will identify key issues that contribute to the DBV literature's methodological heterogeneity and that underlie some of the ongoing disagreements about the appropriate use of meta-analysis in evaluating the DBV. Second, based on our analysis of these issues, we will recommend options for ensuring that meta-analytic investigations address the DBV in ways that are both methodologically sound and

statistically appropriate, and that allow for the DBV hypothesis to be tested fairly.

In order to accomplish these goals, we begin by providing a methodologically focused review of meta-analytic studies of the DBV. Our review includes DBV meta-analyses that address the question of psychotherapy equivalence across all forms of psychological disorders as well as those examining possible equivalence within specific disorders. Although we focus on studies of psychotherapies for adults, many of the methodological issues associated with these studies are also of relevance to the treatment of youth. We begin by reviewing meta-analyses that have included treatments for a wide range of mental disorders in their analyses, starting with Wampold et al.'s (1997) study, which has often been viewed as the *de facto* final word on the topic, and then the four meta-analyses that have been undertaken on the DBV across a range of mental disorders since this study. Following this, we review five disorder-specific meta-analyses that represent a cross-section of several major disorders, methodologies, perspectives, and research groups that feature in the DBV literature. These studies include Siev and Chambless (2007) as an example of meta-analyses for treatments for anxiety disorders; Cuijpers, van Stratten, Andersson, and van Oppen (2008) as an example of meta-analyses for treatments for depression; Bisson, Roberts, Andrew, Cooper, and Lewis (2013), as illustrative of meta-analyses for treatments of trauma; and Baardseth et al. (2013) and Tolin (2014, 2015) which present a sequence of meta-analyses for the treatment of anxiety disorders from critics and proponents of the DBV. This presentation of meta-analytic studies does not constitute a complete review of all disorder-specific meta-analytic studies that have addressed the DBV. However, our selection provides an illustrative review of methodological issues in order to illustrate how to best conduct meta-analyses in order to address longstanding disagreements on DBV research.

The Outcome Olympics: A Critical Review of DBV Meta-Analyses

Wampold et al. (1997)

The meta-analysis by Wampold et al. (1997) is one of the most widely-cited DBV studies of the past twenty years, and it represented a substantial increase in statistical sophistication relative to early meta-analyses on this topic (e.g., Smith, Glass, & Miller, 1980). Wampold et al. (1997) introduced specific methodological controls in their analyses in order to improve the internal validity of their results. For instance, the authors only included studies with head-to-head comparisons of psychotherapies published in high quality journals, thereby controlling for a number of extraneous factors that can arise when treatments are compared on the basis of their effect sizes relative to control conditions (e.g., wait list controls, attentional controls). They also limited their analysis to *bona fide* psychotherapies—defined as those offered by genuine psychotherapists (i.e., individuals with at least a master's degree who made active treatment decisions), geared towards a broadly defined clinical problem, and meeting at least two of the following criteria: citing an established psychotherapeutic approach, citing psychological principles in their rationale, being manualized or codified in a professional book, or having active therapeutic ingredients identified and cited.

Whereas previous meta-analyses (e.g., Smith et al., 1980) had compared different classes of treatments, Wampold et al. (1997) examined the distribution of *all* treatments' randomly-signed effect sizes around zero. That is, rather than having a single reference class to which all other classes of therapy were compared, the effect sizes for each of the comparisons between two treatments were randomly assigned either a positive or negative sign. This produced a distribution of effect sizes centred around zero, the homogeneity of which the authors then tested—tight clustering around zero was taken to indicate that treatment differences generally did

not differ substantially from zero, whereas a heterogenous spread was taken to indicate that treatments differed by a correspondingly large degree. The authors reported no evidence of heterogeneity, indicative of support for the DBV. However, this analytic method is potentially problematic, as most of the comparisons included in the analysis were comparisons of one type of cognitive-behavioural treatment (CBT) to another type of CBT (Crits-Christoph, 1997; Hunsley & DiGiulio, 2002), and it is arguably the case that the critical test of the DBV is undertaken only by evaluating differences between conceptually-*dissimilar* treatment classes.

Rather than directly compare different treatment classes, the authors presented six experts with multiple pairs of treatments and asked them to rate the similarity of the treatments in each pair. Using meta-regression, the similarity ratings of treatment pairs and their associated between-treatment effect sizes were examined, and no relation was found between them. The mean similarity rating across studies was 3.47 on a 7-point scale (mean *SD* = 1.17, ranging from 1.17 to 6.33), indicating that the ratings of the six experts were fairly evenly distributed across the full range of similarity ratings. As the majority of studies were CBT-CBT comparisons, one would expect the mean similarity rating to be much higher if the comparisons captured *absolute* therapy similarity in a way that tracked onto therapy classes. Instead, it may be that the experts' similarity ratings within each pair were anchored to the other pairs of treatments in the set, rather than to an absolute sense of therapy similarity; that is, this indicator of similarity was possibly mapping onto the treatments' similarity *relative to the other treatments in the sample*, a far less-informative measure of similarity or treatment class. Moreover, it's likely that a rating of pure similarity would map poorly onto the conceptual divides that characterize different therapy classes (Crits-Christoph, 1997). Experts would be likely to rate, say, a treatment focused on cognitive restructuring as being quite different from a treatment that emphasized flooding, even

though both fall conceptually under the umbrella of CBT. Our view is that the relative success of these different conceptual frameworks is the true test of the DBV as it is typically presented, but this distinction has not been consistently made or tested in the treatment outcome literature.

By focusing on comparisons between *bona fide* treatments, delivered by suitably trained therapists, Wampold et al. (1997) introduced criteria for study inclusion that were appropriately tighter than was the case with previous meta-analyses in this domain. Nonetheless, their inclusion criteria permitted tests of psychotherapy equivalence for problems such as high school underachievement, social skills training, obesity, and decisional conflict. Arguably an evaluation of *bona fide* psychotherapies should occur with respect to the treatment of *bona fide* clinical disorders, but this perspective is tied up with the contentious question of how best to define psychotherapy. Without unambiguously defining the nature of the test of the DBV, it is likely that analyses may be confounded by assumptions about the proper boundaries of the DBV.

Tolin (2010) and Baardseth et al. (2013)

Rather than employ Wampold et al.'s (1997) technique of exploring effect sizes without categorization into therapy classes, Tolin (2010) took nearly the opposite approach by focusing on one therapy class, CBT, and comparing all other treatments (i.e., psychodynamic therapies, interpersonal therapy, supportive therapies, and "other" therapies) to it for a variety of different psychological conditions. He reported significantly greater effects for CBT overall when compared to the other classes at post-therapy, $d = 0.22$ (95% CI [0.09, 0.35]), six-month follow-up, $d = 0.47$ (95% CI [0.29, 0.66]), and 1-year follow-up, $d = 0.34$ (95% CI [0.06, 0.62]). Subgroup analysis revealed significantly greater effects of CBT at post-therapy relative to psychodynamic therapies and "other" therapies, but not relative to interpersonal or supportive therapies; analyses also indicated that the outcome with CBT was substantially greater than with

psychodynamic therapies at six-month and one-year follow-up. When breaking down the post-therapy results by disorder, he found that CBT outcomes were superior to outcomes for all other therapy classes for depression and anxiety disorders, but not for the other disorders examined in the study (including eating disorders, substance use disorders, psychotic disorders, developmental disorders, habit disorders, personality disorders, marital distress, and academic problems). Moderator analyses found that the rating of the principal investigator's allegiance to a given therapy class significantly predicted study effect size, but that ratings of methodological quality did not. Importantly, potential moderator variables such as these are not always included in meta-analyses evaluating the DBV.

By categorizing therapies into classes, Tolin (2010) addressed some of the concerns about leaving therapies undifferentiated when testing the DBV (i.e., the effects of within-class comparisons outweighing between-class comparisons and the DBV's central test being the comparison of conceptually dissimilar treatments). However, other decisions made in this meta-analysis were less methodologically sound. Many potentially relevant studies were not considered for inclusion in the analyses, as the database search used to identify studies was based on a few keywords related to CBT, combined with a few terms related to non-CBT therapies using the Boolean "AND" operator. Additionally, by conducting moderator analyses despite the lack of significant heterogeneity in the data set, the likelihood of Type I errors was increased.

Tolin's (2010) analyses were critiqued by Baardseth et al. (2013), who reanalyzed his data and presented their own meta-analysis to test the superiority of CBT in anxiety disorders only, which we review in a later section. They criticized Tolin's decision to test for disorder-specific treatment differences by using disorder classes (e.g., testing differential treatments effects for the class of "anxiety disorders," rather than testing each particular anxiety disorder),

arguing that different disorders within a class can have different individual mechanisms of action, and that CBT may be superior in targeting the mechanisms of only some of these disorders, but not others—variability that would be obscured by collecting disorders into classes. Although this point is well-taken, the statistical power required to examine every treatment effect of every disorder would be substantial, rendering most such comparisons prohibitively impractical to conduct.

In their re-analysis, Baardseth et al. separated symptom-specific from more global measures. They reported essentially the same findings as Tolin (2010), except they noted that, for non-symptom specific measures, CBT was not superior overall, nor was it for specific measures of depression; for anxiety disorders, however, the superiority remained even in global measures. This re-analysis underscores the potential for differential findings between symptom-specific and global measures, and hence the value of accounting for both in DBV meta-analyses.

Marcus et al. (2014)

Marcus et al. (2014) chose to follow Wampold et al.'s (1997) approach in using a heterogeneity test to assess the overall DBV, but modified it by analyzing the primary, symptom-specific measures separately from the secondary measures included in a study. They also only included treatment studies that had participants drawn from clinical populations. In contrast to the results found by Wampold et al. (1997), Marcus et al. (2014) found significant heterogeneity for primary symptom measures at both termination, $Q(49) = 107.48$, $p < .001$, $I^2 = 54.41\%$, and at follow-up, $Q(38) = 63.02$, $p = .007$, $I^2 = 39.70\%$. No significant between-treatment heterogeneity was found for secondary measures at either time point. In a supplementary analysis, these researchers found that, at termination, the average absolute effect size difference among treatments for primary measures was $d = 0.29$ (no confidence intervals reported). They

also noted that this difference at pre-treatment was $d = 0.09$ when it should have been zero. They suggested, therefore, the treatment effect size may have been closer to $d = 0.20$ than to 0.29. Using Monte Carlo methods, they determined that this effect size difference was statistically significant at the $p < .001$ level.

The authors also followed the strategy used by Tolin (2010) and Baardseth et al. (2013) in undertaking a contrast of CBT with other therapies (known as the cognitive contrast), finding a superiority for CBT on primary measures at termination, $d = 0.16$, $p < .001$, 95% CI [0.07, 0.26], and at follow-up, $d = 0.16$, $p = .002$, 95% CI [0.06, 0.26]. Consistent with earlier analyses, no CBT superiority was found at either time point on the secondary measures.

This study has received considerable attention from researchers interested in the evaluation of the DBV. Wampold et al. (2017) argued that the validity of the finding for primary measures is undercut by the presence of significant heterogeneity in the data set. However, this would not necessarily undermine the finding of superiority for CBT on primary measures; rather, it is likely to reflect that some studies showed this superiority quite strongly, whereas others showed it quite weakly or not at all. Wampold et al. (2017) also criticized the inclusion of a study by Clark et al. (2006), which they described as biased and as using a flawed protocol for its applied relaxation condition. However, Marcus et al. undertook analyses for their heterogeneity test and for their test of the cognitive contrast that revealed no one study disproportionately affected their findings. Although this demonstrates the advantages of undertaking outlier sensitivity analyses, Wampold et al.' critique about the omission of ratings of researcher allegiance warrants attention, as previous meta-analyses have found a significant effect for allegiance (e.g., Tolin, 2010).

In their critique, Leichsenring et al. (2017) suggested that the Marcus et al. meta-analysis

included data from primary studies that did not meet the criteria for *bona fide* psychotherapy trials (i.e., some studies had insufficient amounts of therapy, therapists were not trained specifically for the treatments evaluated in the study, therapists did not use a manual for the provision of treatment). However, according to the criteria for *bona fide* therapies set out by Wampold et al. (1997), even studies such as these do qualify as tests of *bona fide* therapies. Nonetheless, this criticism suggests that there is not universal agreement on the criteria that should be used to designate a treatment as being *bona fide*, and that DBV researchers may wish to tighten the standard criteria to ensure that studies using weak or inadequate treatments are excluded. Leichsenring et al. also argued that Marcus et al. (2014) failed to include some primary studies of *bona fide* psychodynamic psychotherapy in their meta-analysis. This was likely due to the authors identifying studies for possible inclusion by examining only several key journals that publish psychotherapy research. As the authors used the same criteria used by Wampold et al. (1997), this highlights a recurring issue in the meta-analytic literature on the DBV: there is no widely endorsed strategy for searching for and identifying comparative treatment outcome studies.

Disorder-Specific Trials

The meta-analyses reviewed thus far focused on the overarching question, "Do psychotherapies *generally* differ in their efficacy?" Many additional meta-analyses have examined the question, "Do psychotherapies differ in their efficacy *for a particular disorder or disorder class*?" As the number of disorder-specific meta-analyses vastly outnumbers the meta-analyses that examine the broad question of general psychotherapy equivalence and are far too numerous to comprehensively review here we present a small selection of major, disorder-specific meta-analyses in order to highlight key methodological issues.

Siev and Chambless (2007)

Drawing on the tradition of analyses of *bona fide* psychological disorders, Siev and Chambless (2007) undertook a meta-analysis of five randomized, direct comparisons of CBT versus relaxation therapy for generalized anxiety disorder (GAD) and five such comparisons for panic disorder. The studies were chosen so that the CBT interventions did not have components that overlapped with relaxation therapy. For GAD, no significant differences were found between the treatments, but there were substantial differences found between the therapies on primary measures for panic disorder: $g = 0.34$, $p = .02$ for panic symptoms, and $g = 0.48$, $p = .01$, for panic-related cognitions. No significant differences were found for secondary measures in the panic disorder studies. These results suggest that the DBV may have differential applicability in different disorders. However, as these analyses only examined two particular types of therapy, and as relaxation is often incorporated as a component of CBT in the treatment of anxiety disorders, these results do not address the question of whether differences occur more generally among treatment classes. The small number of studies may have led to underpowered analyses of heterogeneity and, as suggested by Wampold, Imel, and Miller (2009), the results of a single study may have greatly influenced the finding of significant differences for panic disorder.

Cuijpers et al. (2008)

Cuijpers et al. (2008) undertook a meta-analysis of randomized trials of direct comparisons between psychological treatments for depression. Treatments were categorized into CBT, non-directive supportive therapy, behavioural activation, psychodynamic therapies, problem-solving therapy, interpersonal therapy (IPT), and social skills training. Separate summary effect sizes were then calculated for each of the categories relative to all other psychological treatments. For example, IPT was found to be statistically superior to other

therapies, $d = 0.20$. Although the authors suggested that their results offered little support either in favour or against the DBV, it might equally be contended that it offers a limited repudiation of the verdict, with results that indicate the relative inferiority of supportive therapy and superiority of IPT.

There are some methodological issues worth noting in the Cuijpers et al. (2008) meta-analysis. They did not limit the inclusion of studies to those that evaluated *bona fide* psychotherapies, and their classification of therapy type was based on their personal expert judgement (rather than according to criteria derived from treatment manuals, professional organizations, or empirical surveys). For instance, CBT was not found to differ significantly from any of the other therapy classes, but treating social skills training and behavioural activation as treatment classes different from CBT may have reduced CBT's effect size relative to all other therapies included in the analyses. Lastly, because approximately 90 analyses were undertaken in this study, without any adjustment for error rate, the likelihood of Type I errors is of concern.

Bisson et al. (2013)

Bisson et al. (2013) examined direct comparisons of treatments for chronic post-traumatic stress disorder (PTSD) and categorized the therapies into four classes: trauma-focused CBT (TFCBT), standard CBT; eye movement desensitization and reprocessing (EMDR), and a category of "other" psychotherapies. Although the authors reported many outcomes with no evidence of differences among the treatment categories, a considerable number of outcomes *were* reported with substantial differences. For example, on clinician-rated measures of PTSD symptoms, TFCBT was found to be statistically superior to standard CBT at follow-up ($d = 0.51$) and was found to be statistically superior to "other" therapies for self-rated measures of PTSD

symptoms post-treatment (d = 0.60) and at follow-up (d = 0.29). EMDR, although comparable to standard CBT for self-rated PTSD symptoms post-therapy, was superior to CBT at follow-up (d = 0.52); EMDR was also superior to "other" therapies for self-rated PTSD symptoms post-treatment (d = 0.84). Although there are many methodological strengths to this study, it should be noted that the "other" therapies class was heterogeneous in nature, including supportive therapy, non-directive counselling, psychodynamic therapy, and present-centred therapy. Not all therapies included in this class may have met criteria to be considered *bona fide*, and the mix of treatments may have resulted in a mean effect size that does not accurately represent any of the included treatments.

Baardseth et al. (2013) and Tolin (2014, 2015)

As we described previously, Baardseth et al. (2013)—in their critique of Tolin's (2010) study—re-analyzed Tolin's meta-analysis of overall psychotherapy efficacy. They then undertook their own meta-analysis to test the superiority of CBT for the treatment of anxiety disorders. An innovative element in this analysis was that the authors solicited CBT experts listed with the Association for Behavioral and Cognitive Therapies for their opinions as to whether a given treatment constituted CBT for anxiety disorders. Their meta-analysis contained 13 studies of CBT comparisons in the treatment of anxiety disorders, versus four included by Tolin (2010) in his analyses, and did not find superiority for CBT, g = 0.14, 95% CI [−0.08, 0.35]. In Tolin's (2014, 2015) re-analysis of these same meta-analytic data, he found essentially the same results, g = 0.14, p = .21, thereby supporting Baardseth et al.'s contention that the DBV holds in the treatment of anxiety disorders. An important methodological issue raised in by Tolin (2014) is the utility of differentiating between intent-to-treat (ITT) analyses and completer analyses and analyzing them separately. This is a worthwhile strategy which has the advantage

of potentially reducing heterogeneity that might otherwise obscure relevant treatment differences.

Drawing a New Starting Line

The race to establish or overturn the DBV's legitimacy has produced a variety of meta-analyses, with variability in their methodological and statistical approaches that may be contributing greatly to the conflicting conclusions drawn across studies. The meta-analytic literature is sorely in need of tests of the DBV that base their decisions on solid methodological grounds—in other words, tests that are less like the helter-skelter Caucus-race and more like genuine contests with systematic rules and requirements.

We sought to outline the ongoing empirical debate about psychotherapy equivalence by highlighting some of the methodological variability that might have influenced the obtained results. We now turn to presenting a number of methodological "rules" that might be worth adopting and describing our reasons for recommending these rules. In view of the methodological variability that seems to characterize the literature, these are steps we believe are critical in arriving at a truly empirically informed answer to the DBV debate. In the following section we consider methodological elements that have characterized past psychotherapy meta-analyses and underline which options are likely to make for clear and effective tests of the DBV. Many of these recommendations echo those made elsewhere in the DBV literature and some recommendations may seem obvious given the standards for meta-analyses in other fields of health research. It is our hope, however, that consolidating these recommendations will be valuable for psychotherapy researchers and that the inconsistencies highlighted in our review substantiate the need to focus on these particular methodological elements.

Analyses of What?

Direct vs. Indirect Comparisons

The meta-analyses we reviewed all analyzed head-to-head comparisons reported in the primary studies. Generally speaking, such direct comparisons have two major dimensions of superiority relative to indirect comparisons. First, they avoid the problem of non-equivalent control groups. In many indirect studies of psychotherapy effects, treatment as usual (TAU) is understood to function as a control condition, purportedly allowing the authors to demonstrate whether the psychotherapy in question is better than treatment "as usual." However, the norms for TAU vary substantially from study to study, from treatment centre to treatment centre, and over time (e.g., Wampold et al., 2011), meaning that TAUs cannot be considered equivalent comparator conditions across studies. Second, direct comparisons between treatments remove the need for statistical controls of various confounds that might otherwise present a threat to the validity of the comparative data (e.g., confounds by setting, by treatment implementation, or by measurement reactivity; Wampold et al., 1997). In direct comparisons, both therapies under consideration are conducted in the same setting(s), with the same measurements and the same quality of implementation; this may not be the case for indirect comparisons. It is worth noting that indirect comparisons relative to a wait-list control—unlike those relative to TAU—are not plagued by issues of control non-equivalence, as the experience of waiting for treatment is unlikely to vary systematically from setting to setting. However, other confounds (such as different referral sources or recruitment strategies across settings) may still present a problem, again making direct comparisons the first choice for a DBV comparison.

In the past decade, substantial progress has been made in the development of network meta-analyses in which treatment effects are calculated not only directly (i.e., summarizing

studies that compare treatments A and B directly), but also indirectly (i.e., calculating the effect of treatment A versus treatment B *vis-à-vis* a shared comparator treatment C). The benefits and difficulties associated with network meta-analyses have been discussed in detail elsewhere (e.g., Cipriani, Higgins, Geddes, & Salanti, 2013). In brief, network meta-analyses can use indirect comparison effects to vastly increase the statistical power available for meta-analysis and can estimate the relative effects of two treatments that have never been directly compared in a treatment study. On the downside, however, indirect comparisons require an assumption of transitivity: that is, to calculate AB comparisons from AC and BC comparisons, treatment C must be consistent in its characteristics across studies. In psychotherapy research, a given comparison treatment condition often varies substantially from study to study, thus greatly complicating the calculation of indirect effects. Although network meta-analyses will doubtlessly serve as a valuable tool in future DBV studies, they must be undertaken cautiously to preserve validity under the assumption of transitivity. At present, the use of meta-analytic comparisons that rely on direct evidence provides a more straightforward approach to evaluating the DBV.

Bona Fide vs. Intent-To-Fail Treatments.

The meta-analyses we reviewed consisted of analyses of *bona fide* treatments, almost all using Wampold et al.'s (1997) criteria. These criteria are a minimum requirement for ensuring that psychotherapies being compared are true, real-world "competitors," rather than treatments that are intended to fail so as to prove the viability of another treatment (Wampold et al., 1997) or are composed of few elements that are likely to prove to be therapeutic. However, we agree with Marcus et al.'s (2014) suggestion that it would also be appropriate to include studies in which therapy was administered by psychotherapy trainees under the direct supervision of licensed psychotherapy providers. If such studies were included in DBV meta-analyses, the

licensure status of the therapists (i.e., licensed vs. trainees providing therapy under supervision) could be examined as a potential moderator variable.

What Constitutes a "Clinical Problem"?

We suggest that tests of *bona fide* psychotherapy should be geared towards the treatment of *bona fide* "clinical problems." As discussed earlier, Wampold et al.'s (1997) meta-analysis included tests of therapies for problems such as high school underachievement for which most manualized psychotherapies were not developed. Inclusion of data from psychotherapy trials for these problems may have made it more difficult to detect potential treatment effects among trials conducted in populations for which these therapies were developed—that is, *bona fide* clinical populations, consisting of individuals with problems of a clinical nature and severity. Some more recent DBV meta-analyses have focused only on trials in which individuals were diagnosed with a mental disorder or scored above a clinical cut-off on psychometrically strong measures of psychopathology (e.g., Marcus et al., 2014). This is not to suggest that non-diagnosable problems are not an important clinical or research focus; indeed, limiting analyses to diagnosed mental disorders runs the risk of excluding important conditions from consideration. Nonetheless, such a focus is a straightforward way to reduce a potential source of error in the meta-analytic data. Researchers examining comparative treatment effects might find it fruitful to consider the question of the DBV for non-diagnosable problems as a separate question from that of the DBV for diagnosed disorders.

…Comparing What Outcomes?

Primary and Secondary Outcomes

Some researchers (e.g., Tolin, 2010) have considered the possibility of differential effects between symptom-specific ("primary") and global ("secondary") measures when evaluating

outcomes in meta-analyses. Primary measures have been the dominant outcome measure of concern in most treatment studies and most meta-analyses have focused on results obtained with primary measures. Wampold et al. (2017) argued against the importance that primary measures have been afforded in meta-analytic tests of the DBV. They contended that scores on primary measures may be *correlated* with clinically important outcomes but are not themselves of clinical importance, as they are simply proxies for the true outcome of interest. With the growing research evidence that all mental disorders may be undergirded by a fundamental shared factor of psychopathology (e.g., Caspi et al., 2014), they posited that disorder-specific measures may be measuring outcomes that are merely coincident with this underlying factor. Wampold et al. suggested, therefore, that researchers should use data from secondary measures of distress (e.g., quality of life), which they suggested may better address this factor.

However, it seems misguided to suggest that primary measures are inferior to secondary measures. For example, when treating individuals with obsessive-compulsive disorder, changes to obsessive beliefs and compulsive behaviours are important considerations for the success of treatment. Even if it is true that these compulsions are manifestations of an underlying factor, their reduction would still be a valuable endpoint because of their obtrusive, disturbing character—they simply are not clinically neutral phenomena. Furthermore, it has been acknowledged for decades that there are important conceptual differences across forms of psychotherapy about whether clients' symptoms and problems should be considered as being signs of an underlying condition or as worthy of treatment in their own right (cf. Goldfried & Kent, 1972). In light of these conceptual differences, we argue both primary and secondary measures should be regarded as relevant to the evaluation of treatment in their own rights—to do otherwise would be to give short shrift to one or more conceptualizations of therapy. Consistent

with the recognition of the potentially unique contributions of both of these types of outcome measures to understanding psychological distress, meta-analyses of psychotherapy are increasingly examining whether (a) meaningful differences exist between psychotherapy classes on either type of measure and (b) the pattern of any differences is the same for primary as for secondary measures. We believe that examining these questions should be standard practice in meta-analytic studies of comparative treatment outcome.

Results at Termination and Follow-Up

Psychotherapy outcome trials frequently include data from the period immediately following the termination of therapy and from follow-up periods. Where possible, including analyses of data from both of these periods would enrich meta-analytic tests of the DBV, helping to elucidate whether differences exist at different periods and if treatment gains (and any differential treatment gains) are maintained.

Intent-to-Treat and Completer Analyses

Some researchers have separated data from studies which use completer analyses from those which use intention-to-treat (ITT) analyses (e.g., Tolin, 2010, 2014). In the former strategy, analyses are undertaken only on the data of participants who completed the treatment. In ITT, by contrast, analyses are undertaken on all participant data regardless of what occurred after the randomization procedure; missing data are treated using such procedures as multiple imputation or carrying forward the last data value provided by participants. ITT analyses conserve randomization and hence have reduced bias relative to completer analyses. As they also directly account for the effects of the clinically relevant issues of noncompliance and premature termination, ITT analyses are recommended in major reporting guidelines for clinical trials (e.g., Schultz, Altman, & Moher, 2010). However, relative to completer analyses, ITT analyses are

conservative, as they are likely to underestimate treatment differences (Gupta, 2011). For this reason, meta-analytic researchers should undertake sensitivity analyses to establish a plausible range for the summary effect when comparing completer to ITT analyses (Higgins & Green, 2011).

Effect Size Interpretation

Once a meta-analysis has been undertaken and a summary effect value is produced, the question arises as to what magnitude of between-treatment effect should be considered clinically significant, or as substantial enough to be taken as falsifying the DBV. Wampold and his colleagues have, over the past two decades, repeatedly argued that an effect size of approximately $d = 0.20$ should be considered small and clinically insignificant in the context of the DBV (e.g., Messer & Wampold, 2002; Wampold et al., 1997). However, this is contentious: an effect size of this magnitude has also been considered to provide meaningful evidence against the DBV (Marcus et al., 2014; Tolin, 2010), and even Benish, Quintana, and Wampold (2011) have argued that such an effect size signals the superiority some psychotherapy adaptations relative to others.

What should, then, be considered clinically meaningful when considering comparisons between treatments? Although commonly used rules of thumb treat an effect size of $d = 0.20$ as small, these were proposed simply as rough heuristic devices (Cohen, 1988). A more appropriate response to the question of what effect size should be considered clinically relevant is that it depends on context, including the nature of the effect under consideration and the nature of the data from which the effect size was derived (Hunsley & Westmacott, 2007). For example, effects for changes in central characteristics of a disorder should be viewed differently from changes in more peripheral characteristics and effects derived from experimental research designs should be

viewed differently from those derived with correlational designs.

A central consideration for the relevance of a between-treatment effect size is what such a difference would mean for a client. One way this might be gauged is to consider what effect size would correspond to a change that clients could themselves subjectively detect (i.e., the "minimally important difference" [MID] approach). Cuijpers, Turner, et al. (2014) undertook preliminary research into just this question, finding that an MID of roughly $d = 0.24$ or greater would be equivalent to a noticeable change in mental well-being, although further research is required to ascertain the stability of this value across measures and to determine MIDs for a number of different domains and populations.

Alternatively, an effect size can be converted to a metric that demonstrates the strength of the effect *across* participants/clients—that is, its strength within the population receiving treatment. The number-needed-to-treat (NNT) is such a metric and provides an estimate of the number of people that one would need to be treated with therapy A in order to have one more successful outcome than would occur with therapy B. Simple conversion formulae are available to calculate NNT from various types of effect sizes (e.g., Kraemer & Kupfer, 2006). Using such a formula, a value of $d = 0.20$ yields an NNT of 8.9: this indicates that if therapy A has outcomes better than therapy B at $d = 0.20$, then one person would be likely to have better relative outcomes with A for approximately every 18 people who were treated (i.e., if 50% were treated with A and 50% with B; Furukawa & Leucht, 2011). On the scale at which psychotherapy is implemented in health care systems, this would translate to tens of thousands of individuals who would receive improved outcomes if given A rather than B. The NNT may therefore offer a fairly intuitive way of reframing the standardized mean difference, expressing treatment differences clearly in terms of their impact on clients. We recommend, therefore, that when

interpreting effect sizes in research relevant to the DBV, researchers carefully consider the clinical implications of the effect sizes they obtained. This should provide a far more meaningful interpretation of study results than is obtained by relying on simple guidelines that are insensitive to clinical context.

…Undertaken with What Methods?

Search Strategies

A number of different strategies have been undertaken in the DBV literature to locate primary studies for meta-analysis. Wampold et al. (1997) and Marcus et al. (2014) undertook their analyses using an index search of several key journals known for publishing high-quality studies of psychotherapy outcomes. Cuijpers et al. (2008) undertook a database search to find psychotherapy outcome studies for a specific disorder (i.e., depression), whereas Tolin (2010) undertook a database search for all psychotherapy outcome studies regardless of clinical population. Finally, Bisson et al. (2013) searched the Cochrane Common Mental Disorders Control Trials Register (CCMDCTR), regularly populated from other databases using comprehensive search strategies.

Two major concerns when selecting a search strategy are sensitivity (i.e., the exhaustiveness of the search and its representativeness of the literature) and specificity (i.e., the ability of the search to avoid capturing studies unrelated to the aim of the meta-analysis). The aim is to capture as many relevant studies as possible without making it prohibitively difficult to subsequently exclude irrelevant studies. The Cuijpers et al. (2008) strategy is a strong option, as limiting the search to a single disorder allows for a comprehensive search for psychotherapy studies without being encumbered by irrelevant studies. However, such a strategy doesn't allow for a comprehensive test of the DBV—only a disorder-specific one. The related strategy used by

Bisson et al. (2013) is an excellent option, as CCMDCTR aims to centralize all (and only) studies relevant to controlled trials of mental health treatment outcomes; however, access to the CCMDCTR is not available for researchers who are not members of the Cochrane Common Mental Disorders group. The strategy used by Wampold et al. (1997) and Marcus et al. (2014) studies is specific, but it may be insufficiently sensitive. Furthermore, by limiting their searches to top journals, these authors' studies may have been influenced by potential biases on the part of those journals (Leichsenring et al., 2017). Tolin's (2010) search was likely not encumbered by many irrelevant studies because his search terms were not exhaustive—rather than using a combination of key words and subject/MeSH terms, he used a few simple key words which were likely too specific (e.g., the term "psychotherapy" was not used). However, if this search had been more sensitive, it would likely have resulted in a very high number of irrelevant studies.

In sum, when choosing a primary search strategy for disorder-specific DBV meta-analyses, comprehensive database searches are an optimal choice, allowing for sensitive and comprehensive searches that are simultaneously manageable due to the disorder-limited scope. For non-disorder specific DBV meta-analyses, however, because of the vast number of irrelevant studies likely to be included in a comprehensive database search, there is no clearly superior primary search option. Both journal index searches and insensitive database searches compromise sensitivity and may bias the meta-analysis, but such compromises may be necessary for the sake of feasibility.

Swift and Wampold (2018) recently offered a detailed discussion of key considerations for the study inclusion and exclusion process in psychotherapy meta-analyses, including a discussion of search strategies. In addition to using search terms to search general databases, seeking out specialized databases, and hand-searching journal indexes, the authors also

suggested root-and branch searches of relevant articles, reviews of previous meta-analyses, and discussions with experts in the field. Importantly, the authors recommended combining multiple search strategies, a suggestion with which we concur—this way, deficits in one strategy may be addressed by using another. In particular, if researchers intend to meta-analyze more than the most recent studies, they should review the reference lists of relevant earlier meta-analyses to ensure that they do not miss previously included primary studies.

Two additional search-related considerations are worth highlighting. First, researchers should describe how they conducted the searches in full detail, including the search terms and limits used, so that other researchers can replicate their findings. Second, researchers should make efforts to search the so-called "grey literature" (i.e., sources of unpublished studies, including unpublished dissertation studies) so as to counteract possible concerns related to publication bias (see "Statistical Considerations," below). As not all grey literature is of equal quality, researchers should be selective with the unpublished literature upon which they choose to rely. Indeed, when using studies from the grey literature, moderator analysis may become especially important for determining whether studies' methodological quality impacts their reported effect sizes (see "Moderator Analysis," below).

Statistical Considerations

Standards for meta-analysis have changed substantially since the DBV was first addressed meta-analytically (Smith et al., 1980), and there are several state-of-the-science statistical considerations that should constitute a minimal requirement for all meta-analyses of treatment outcome. When synthesizing the summary effect, effect sizes from individual studies should be weighted by their precision (i.e., weighted proportionally to their sample size). Because studies of therapy outcomes are drawn from multiple populations (i.e., there is assumed

to be genuine between-study variance), a random-effects model should be used if the number of included studies, k, is sufficiently high (e.g., $k \geq 10$; Higgins & Green, 2011). The meta-analysis should include sensitivity analyses to assess the impact of outlying effect sizes as well as of outlying sample sizes. Heterogeneity of the summary effect should be assessed using multiple measurements, including Cochran's (1950) Q test to assess for the presence of heterogeneity, as well as the estimated variance of the true effects (T) and the prediction interval ($\cong 4T^2$) to evaluate the degree of spread of the true effects. Although I^2—an indication of the percentage of observed heterogeneity attributable to variation in true effects—has historically been more popular as a means of quantifying heterogeneity, the prediction interval is increasingly preferred, as it indicates in absolute terms how much the effect size is expected to vary from one study to another (Borenstein et al., 2009). The Q test should be assessed at $\alpha = .10$ due to its low sensitivity to detect heterogeneity (Fleiss, 1981). If substantial heterogeneity is indicated, moderator analysis should be undertaken (see below). This should include the calculation of the meta-analytic R^2 analogue and a goodness-of-fit test (e.g., the test of $Q_{residual}$).

To account for the possibility of publication bias, several analyses should be included, such as funnel plots and Egger's linear regression test (Egger, Smith, Schneider, & Minder, 1997), Orwin's fail-safe N (Orwin, 1983), and trim-and-fill procedures. Testing for publication bias allows for an assessment of whether the effects obtained in the analysis can confidently be ascribed to true treatment differences, as opposed to having drawn upon a skewed initial pool of studies. Lastly, if multiple dependent measures are used in a study (e.g., multiple measures of depression), these should not be treated as independent data points in a single meta-analysis; rather, a synthetic combined effect should be meta-analyzed, produced using the mean effect and a pooled variance estimate based on published values of the correlation between the two

measures (Borenstein et al., 2009).

Overall Comparisons, Pairwise Comparisons, and Omnibus Tests. In testing the DBV, there are two questions of ultimate concern. First, are there overall differences in efficacy among psychotherapy classes and, if there are differences, what is their magnitude? This is the test of the DBV in its strictest sense. Second, if there are such overall differences, which of the treatment(s) demonstrate superiority and by how much? A standard meta-analytic strategy is to categorize the treatments into categories (e.g., CBT, psychodynamic, IPT), and meta-analyze the effects of each treatment type relative to all others, testing to see if any category has a summary effect size that is significantly superior to the rest. This method answers the second part of the question directly, thereby indirectly answering the first. However, this analytic strategy does not control for issues of family-wise error rate. Accordingly, it would be preferable to conduct an overall test of the DBV as an omnibus test before undertaking subsequent pairwise comparisons.

One method to accomplish this is to produce a summary effect size from the absolute values of each pairwise comparison of post-treatment effects, as this represents the weighted average difference in efficacy across psychotherapies. If a significant overall effect exists, and appropriate tests (e.g., the *Q* test) indicate substantial heterogeneity among the effects, then the second portion of the key question could be answered through examination of the potential moderating effects of therapy class (among other moderators; see below). This sort of aggregation of effects is considered to be the standard meta-analytic procedure (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009). However, aggregating the absolute values capitalizes on chance: initial differences between groups due to error—which might normally occur randomly in the positive and negative direction—are all given a positive sign and summed, leading to overestimated final group differences. Although far from a perfect solution, one can follow the

strategy used by Marcus et al. (2014) and subtract the weighted average of pre-treatment group differences from the weighted average of post-treatment group differences, producing a rough estimate of the overall DBV effect. Alternatively, as outlined by Marcus et al., Monte Carlo methods can be used to account more precisely for the issue of capitalization on chance.

A second method that has been used (primarily by Wampold and his colleagues) is to allocate the signs of treatment such that, under the null hypothesis, one would expect a homogenous distribution of effect sizes—for example, by randomly distributing the signs so they are distributed homogenously about zero—and then testing for heterogeneity in this distribution. If the effects are closely clustered around zero, this supports the null hypothesis of no overall differences, whereas a large spread of effects would not support the null (Wampold et al., 1997; Wampold & Serlin, 2014). This method would serve as an omnibus test of overall treatment differences, thereby justifying further pairwise treatment differences without worry of inflated error rates. However, because random distribution of signs necessarily produces an effect size of approximately zero, this method precludes the production of an overall summary effect size, and hence precludes moderator analysis of that summary effect. Furthermore, like the standard meta-analytic approach, this method also capitalizes on chance. As Tran and Gregor (2016) pointed out, random distribution of signs only approximates zero as k approaches infinity and likely does not equate to zero for small values of k; however, the test of homogeneity assumes a null average effect. For example, Benish, Imel, and Wampold (2008) reported nonsignificant heterogeneity for their meta-analysis of treatment differences for PTSD, but when Tran and Gregor (2016) applied Monte Carlo methods to the same studies used in this meta-analysis, they found significant heterogeneity 20 out of the 30 times in their analyses. Accordingly, applying Monte Carlo methods in conjunction with the homogeneity test may prove to be the best option for

correcting the issues stemming from chance effects and low statistical power.

Treatment Classification. Whether one uses the standard meta-analytic approach or the homogeneity-testing approach, we believe that researchers should carefully consider the merits of classifying the individual treatments from each study into their overarching classes (e.g., CBT, psychodynamic, IPT). Two potential benefits of this process are particularly worthy of consideration. First, it allows researchers to assess whether the overall effect differs between studies that compare treatments within a class versus those that assess treatments between classes. Second, indexing treatments by their classes allows researchers to more easily undertake pairwise contrasts when trying to answer the second component of the DBV question (i.e., which treatments, if any, are superior). The method most frequently undertaken here is the so-called "cognitive contrast" (Tolin, 2010), wherein a meta-analysis of CBT-versus-non-CBT comparisons is undertaken, using CBT as the reference therapy class to which the others are compared. Choosing CBT as the reference therapy class has previously been undertaken specifically as a test of the claim that CBT is superior to other treatments (e.g., Baardseth, 2013; Marcus et al., 2014; Tolin, 2010). However, there are more utilitarian reasons for undertaking it—namely, that CBT is the most frequently studied type of therapy, that most between-class, head-to-head psychotherapy studies will be comparisons of CBT to another class of psychotherapy, and that the cognitive contrast will provide the most statistical power of any of the potential between-class comparisons. It is these issues of statistical power which we believe lend the cognitive contrast the most weight as an analytical strategy.

None of this, however, is to suggest that the classification of therapies into categories is a straightforward prospect. Indeed, of all the methodological dimensions upon which DBV meta-analyses differ, treatment classification differences may well be responsible for much of the

observed differences among meta-analytics studies (cf. Baardseth et al., 2013, Tolin, 2014). The process of classifying treatments into categories has been criticized as being arbitrary and significant variation have been noted from one meta-analysis to another (e.g., Wampold et al., 1997; Wampold et al., 2017). Indeed, researchers have been enjoined to recognize the boundaries of psychotherapy classes are not "real"—they are arbitrated not by nature, but by social consensus and dispute (Wampold et al., 2017). Although the benefits of classification are significant, it is unlikely that any one classificatory scheme will be fully satisfactory to all researchers and clinicians. However, the fact that therapy classes have social rather than natural boundaries may offer one avenue for a reasonably acceptable solution—namely, to classify on the basis of how therapy categories are employed by the community of psychotherapists. This might involve a Delphi study, in which experts in the field are iteratively consulted until they converge upon definitions for each category. To our knowledge, no such study has yet been undertaken for psychotherapy classes. In lieu of this, researchers might derive characteristics of therapy classes from treatment manuals or professional organizations representative of each class. The empirical method used by Baardseth et al. (2013)—obtaining similarity ratings from therapy experts—also provides a useful avenue for generating relatively widely agreed-upon characteristics of psychotherapy from the psychotherapy community itself. Such ratings, though, should be based on broad considerations about the theoretical assumptions and treatment principles that differ across therapy classes so as to avoid possible misclassifications of treatments based on simple surface features.

Moderator Analysis. If a statistically significant overall effect is found suggesting psychotherapies have differential effects, it is important to assess whether this effect is consistent or discrepant across studies using a test of homogeneity (i.e., whether all studies show the same

effect, or whether some studies show this effect whereas others do not). If the effect is present in only some studies, it is important to further assess for potential study variables that might account for the heterogeneity; in other words, testing for moderators of the overall effect. There are a number of important potential moderators to consider in evaluations of the DBV.

Analyses Within vs. Between Treatment Classes. First, there is the matter of assessing whether studies that compare treatments from *the same* class systematically differ from those assessing differences of *between-class* treatments. As discussed above, several meta-analyses involved overall or omnibus analyses in which a substantial number of head-to-head comparisons were within-class comparisons (e.g., CBT-versus-CBT comparisons). Because the DBV is fundamentally concerned with differences between psychotherapy approaches, this may have had the unfortunate effect of unnecessarily inflating error variance and potentially disguising relevant treatment differences. Assessing whether between- or within-class status moderates the overall effect would help to determine if this hypothesis is, in fact, correct. If supported, it may be worthwhile to meta-analyze between-class comparisons separately from within-class comparisons.

Disorder. A further potential source of heterogeneity might be the disorder in which treatment differences are assessed. Marcus et al. (2014), for instance, found evidence that the DBV may not hold for disorders characterized by symptom-specific distress (e.g., tic disorders, panic), wherein symptom-specific treatments had better outcomes than did less-focused treatments. In contrast, they found that the DBV was generally supported for disorders characterized by more diffuse symptoms, such as depression. Similar patterns were also observed in the Siev and Chambless (2007) meta-analysis of anxiety disorders. Future analyses should therefore account for the role of disorder type as a possible moderator.

Allegiance. Limiting analyses to head-to-head comparisons controls for many confounds, but some remain salient, and the degree to which they arise in the comparisons may explain heterogeneity in the observed effects. One such important consideration is researcher allegiance—that is, whether the research team, particularly the primary investigator, is likely to be invested in the success of one therapy family relative to the others. Past meta-analyses which have included allegiance as a moderator have found that effect sizes can, indeed, be related to the allegiance of the primary researcher (e.g., Luborsky et al., 1999; but see Leykin & DeRubeis, 2009). Some DBV meta-analyses have only found an effect of the allegiance of studies' primary investigators and not for the allegiance of studies' research teams or of the therapists who delivered treatments (Tolin, 2010). Allegiance can be assessed in a number of ways, including reprint analyses (i.e., analyzing publication history), self-ratings, and ratings by colleagues (e.g., Luborsky, 1999). Perhaps the clearest measure of allegiance is whether a researcher developed the treatment under investigation—this should be accounted for by any allegiance measure (Leykin & DeRubeis, 2009). We echo recommendations made repeatedly in the comparative treatment literature to develop research teams with mixed allegiances whenever possible (e.g., Luborsky, 1999; Leichsenring et al., 2017) and to include researcher allegiance as a possible moderator in meta-regression analyses.

Methodological Quality. Finally, it is important to account for methodological quality of the primary studies included in the meta-analysis—that is, whether the observed effect is only found (or is only absent) in poorly conducted studies. Generally, measures of methodological quality serve to quantitatively assess a number of methodological considerations, including randomization, blinding, trial pre-registration, or use of intent-to-treat analyses. A variety of these measures are available, included those based on criteria developed by Foa and Meadows

(1997), Foa, Keane, and Friedman (2000), Cuijpers et al. (2008), Moher, Liberati, Tetzlaff, Altman, and the PRISMA Group (2009), Higgins and Green (2011), and Shea et al. (2017). Meta-analysts should keep these criteria in mind and possibly combine or adapt them for the context of psychotherapy research (e.g., double-blinding is not a realistic criterion for psychotherapy research).

Conclusions and Recommendations

The methodological variety across the DBV's meta-analytic literature can be considerable; however, when considered in their proper context and *in toto*, the strengths and limitations of the various methodological options become clearer. In the end, researchers' precise research questions and critical decision-making will be their ultimate guides when choosing how to conduct meta-analyses. However, on the basis of the issues that arose from our review, we offer a few recommendations to make the DBV literature a little less helter-skelter than Alice's Wonderland. Although these recommendations cannot serve to guarantee the quality of a meta-analysis, they can serve as a starting point for ensuring a minimal level of agreement among researchers.

1. Direct, head-to-head comparisons of active, *bona fide* psychotherapies should be considered the gold standard for meta-analyses of psychotherapy outcome.
2. Psychotherapies should ideally be evaluated in *bona fide* clinical populations.
3. Researchers should use comprehensive database searches to locate studies where possible; where it is not feasible, they should combine multiple other search strategies, using their judgement to balance the sensitivity and specificity of these strategies. Researchers should be as detailed and as transparent as possible in reporting their search strategy, and they should make efforts to include high-quality grey literature such as

unpublished doctoral dissertations.

4. Where possible, analyses should be undertaken of both primary and secondary outcomes, of both post-treatment and follow-up data, and of both ITT and completer data.
5. Classifying treatments into superordinate therapy classes, and then conducting initial analyses at the level of therapy classes, is recommended as a way of extracting the greatest number of useful comparisons from the meta-analysis. However, researchers should be mindful of the conceptual difficulties associated with this process. Categorizing with reference to treatment manuals or professional organizations' standards for the primary characteristics of each therapy class are viable methods to decrease the arbitrariness of classification.
6. Researchers should employ standard statistical methods for state-of-the-science meta-analyses, including random-effects modelling, outlier and publication bias analysis, homogeneity tests, and moderator analyses (when significant heterogeneity is present). Moderators should minimally include: within- or between-therapy-class status, disorder, methodological quality, and researcher allegiance.
7. To control for family-wise error rates, an omnibus test of the overall DBV effect should be undertaken. Both homogeneity-testing methods and standard meta-analytic methods are reasonable approaches to aggregating effect sizes; however, the standard meta-analytic method can produce a summary effect that is available for moderator analysis, and is therefore preferred. In either case, one should account for the possibility of capitalizing on chance initial treatment differences (e.g., with Monte Carlo methods).
8. Unless there is sufficient statistical power available to undertake pairwise comparisons, the relative superiority of different therapy classes should be tested using the "cognitive

contrast" as it is likely to provide the most powerful test of differences.

9. The clinical significance of effect sizes should not be assessed using commonly applied rules of thumb; rather, they should be considered in the context of the nature of the effect, the data contributing to the effect size, and whether such an effect would be expected to translate into reasonably meaningful clinical outcomes for clients.

Given the theoretical and practical stakes of the outcome of these meta-analyses—including their implications for client welfare and professional training priorities—it is paramount that methodological and statistical decisions are geared towards producing accurate and fair tests. Specifically, it is critical to know if a given therapy is the best first-line treatment option available for a client or, conversely, if there are several therapies with comparable empirical support that should be considered. In order to make such determinations, the hypothesis of psychotherapy equivalence must be evaluated in a truly scientific manner, rather than by the running of another Caucus-race.

***Primus Inter Pares*?**

A Meta-Analysis of Psychotherapy Outcome Equivalence for Mental Health Disorders

Shawn G. Sanders

The University of Ottawa

Reference

Sanders, S. G., & Hunsley, J. (2026). *Primus inter pares*? A meta-analysis of psychotherapy outcome equivalence for mental health disorders [Unpublished manuscript]. School of Psychology, University of Ottawa.

Primus Inter Pares? A Meta-Analysis of Psychotherapy Outcome Equivalence for Mental Health Disorders

The Dodo Bird Verdict (DBV)—named for the Dodo bird's proclamation in *Alice's Adventures in Wonderland* that "*everybody* has won, and all must have prizes" (Carroll, 1865/2008, Chap. 3, para. 20)— is the proposition that all psychotherapies are equivalently effective to one another (e.g., Luborsky et al., 1975). The DBV—and whether it truly reflects the relative efficacy of psychotherapy—seems at least superficially to bear on several contentious issues, including the question of whether common or theory-specific change mechanisms drive the efficacy of psychotherapy; the question of whether some theories of psychological functioning (e.g., cognitive behavioural; psychodynamic) are superior to others; and the practical question of how best to alleviate psychological suffering. Unsurprisingly, the DBV has been hotly debated over the past half-century.

Almost immediately after it was proposed as a verdict on the outcome literature (Luborsky et al., 1975), the appraisal of the DBV became closely connected to a particular methodological approach: meta-analysis. Indeed, the first-ever meta-analytic studies were conducted to synthesize the research on psychotherapy outcomes and thereby provide a comprehensive evaluation of the efficacy of psychotherapy (Glass, 1976; Smith & Glass, 1977; Smith et al., 1980). Since these initial studies, thousands of meta-analytic studies have been published assessing the efficacy of psychotherapy, with numbers growing every year. Indeed, a *single* researcher has published over 350 meta-analyses on psychotherapy to date (Cuijpers, 2025). Some of these myriad analyses have suggested equivalent outcomes among treatments, seemingly supporting the DBV. Others appear to imply that although many psychotherapies are effective, certain treatments are more effective than others in some circumstances, apparently

supporting a converse hypothesis to the DBV: that some treatments, like the Archbishop of Canterbury, are *primae inter pares*: firsts among equals (Anglican Communion Office, n.d.).

Nevertheless, definitive consensus on the DBV has not yet been attained. Arguably, this lack of agreement may be attributable to a set of interrelated methodological and conceptual issues; specifically, conceptual vagueness in what is meant by "psychotherapy equivalence" can produce methodological decisions for the meta-analytic tests, which in turn lead those tests to be relevant for one conception of the DBV, but not for others. In other words, different meta-analyses have been testing competing and occasionally incompatible "Dodo Bird Verdicts." Much like the animals in Carroll's *Wonderland*, researchers have inadvertently been running races with different starting points and toward different ends.

The specifics of many of these methodological matters have been reviewed in detail (e.g., Sanders & Hunsley, 2018). To summarize some key issues: most meta-analyses relevant to the DBV contrast different psychotherapies' absolute efficacy relative to a control group such as "treatment-as-usual" (TAU), waitlist, inactive intervention, or a combination of these. However, these control groups are non-equivalent to one another (e.g., Cuijpers, Miguel, Harrer, et al., 2024), even within one class of controls (Wampold et al., 2011); for example, what constitutes TAU may range from brief check-ins with physicians in some studies to *bona fide* cognitive-behavioural therapy (CBT) in others. Taking such indirect comparisons as conceptually inappropriate for making determinations about the DBV is likely to lead researchers to focus their analysis on the relative efficacy of psychotherapies based on only direct, head-to-head comparisons. Related to the notion of "inactive interventions" is the decision to compare *any* psychotherapeutic intervention including control interventions, to limit to "*bona fide*" treatments (i.e., those meant to be "real-world" psychotherapies; e.g., per the criteria of Wampold et al.,

1997). This decision may be rooted in whether one conceives of the DBV to be primarily about testing the incremental benefit of specific ingredients over common factors, in which case a comparison to even so-called "intent-to-fail interventions" may be defensible, or about differences between actual treatments, in which case such interventions may be inappropriate comparators. The decision to limit comparisons to populations with diagnosed mental health disorders, or to include non-clinical problem areas, may depend on how broadly one construes the DBV and on the importance that the researcher places on being able to estimate the impact of the population on the effect size using reliable categories. How the researcher construes the most relevant measurement of the impact of the interventions—immediately post-treatment, or long-term effects—may affect whether they aggregate post-treatment or follow-up effect sizes; similarly, whether they consider the most relevant measure of impact to be disorder-specific or global may affect what sort of outcomes they choose meta-analyze. Whether one believes the DBV refers to equivalency among treatment classes (e.g., cognitive behavioural therapies, humanistic therapies) or among specific interventions (e.g., exposure and response prevention, two-chair technique) is likely to impact whether one aggregates effect sizes using standard meta-analytic techniques (i.e., classify treatments and choose a reference class to determine the sign of the summary effect size) or choose not to classify treatments (e.g., distribute effect size signs randomly and test heterogeneity, and use absolute values to generate the summary effect size; Wampold et al., 1997, Wampold & Serlin, 2014). Lastly, one's views about the DBV may shape which, if any, moderators of effect size one chooses to test.

The aim of this meta-analysis is to stake out a particular conception of the DBV and produce, given that conception and given practical constraints, proceed with methodologically

defensible approaches to meta-analytic testing. To that end, a view of psychotherapy equivalence is adopted for this study such that:

1. "Psychotherapy" is understood as
    a. Superordinate "families" of treatments,
        i. Comprising clusters of delineated treatment "packages" as delivered in individual trials, which are
        ii. Clustered based on whether they demonstrate particular techniques, which are themselves
        iii. Classified into "families" based on their theoretical and historical identity as comprising an "orientation."

    This sort of grouping of interventions into families is intended to capture what, in real-world practice, *therapists actually ascribe to*. Generally speaking, psychologists and psychotherapists generally do not identify with particular therapy techniques but rather with a particular theoretical orientation of therapy (or mix of orientations), which then guides their selection of technique; for example, in a survey of psychotherapy clinicians, 76% of participants identified with a specific theoretical family as their primary orientation (Tasca et al., 2015)[4]. This high-level grouping of treatments is intended to be a concrete analogue of the theoretical approach, and therefore of most practical relevance to working therapists. Using individual techniques to define these high-level groupings makes it straightforward to appraise manuscripts and categorize the

[4] However, in the same survey, "eclectic/integrative" was the second most popular orientation at 18% of respondents, following CBT/BT/DBT at 34%. Moreover, the relation between therapeutic orientation and use of therapeutic techniques in practice is not entirely clear-cut; see Creed et al., 2014, and von Ranson et al., 2012.

treatment packages they describe. However, this decision is also likely contentious, as the extent to which specific techniques are "proprietary" to specific theories is debatable (e.g., Ahn & Wampold, 2001). In this meta-analytic study, a rating scale (McCarthy & Barber, 2009) was used to increase the reliability and reduce the arbitrariness of these taxonomic choices.

   b. Real-world treatments, operationalized as a *bona fide* criterion for interventions in the primary studies to exclude comparisons with "intent-to-fail" treatments (e.g., Wampold et al., 1997).

2. "Equivalence" is viewed as

   a. Most clearly established by randomized controlled trials (RCTs) of direct, head-to-head comparisons between treatments. Accordingly, only these kinds of comparisons are included in this meta-analytic study.

   b. Most relevant when limited to a verdict within problem areas (i.e., psychiatric diagnoses) rather than as a determination about psychotherapy "overall." Accordingly, the effect size for therapies "overall" was calculated as a global test, and moderator analysis was used to appraise the impact of the population in terms of psychiatric disorder.

   c. Established most compellingly by comparing whether the effect size of comparisons *within* a treatment family (e.g., one CBT treatment compared to another) differ from that of comparisons *between* families (e.g., a CBT versus a Psychodynamic treatment). In line with this, whether a study is a within- or between-family comparison was tested as a potential moderator of the effect size (see "Effect Size Relevance," below).

d. Relevantly evaluated based on

    i. (a) Measures of outcome that are specific to symptoms characteristic of the primary problem area, (b) symptom-specific measures more generally, and (c) measures of global functioning. Primary disorder-, secondary disorder-specific, and global measures were all coded and meta-analyzed separately.

    ii. Clients' outcomes in the immediate aftermath of therapy as well as in the long-term. To that end, effects were collected from both the post-treatment and follow-up timepoints and separately meta-analyzed.

    iii. Whether potential sources of bias impact the meta-analytic result. Accordingly, analyses were undertaken to appraise the impact of publication bias, researcher allegiance to particular treatments, and risk of methodological bias (e.g., through moderator analysis).

When Should I Care? The Issue of Effect Size Relevance

In addition to staking out a particular conception of the DBV, in this paper, a particular conception of a relevant effect size will be adopted. The issue of effect size interpretation—that is, what degree of between-treatment effects is of practical significance for making determinations about the DBV—is both important and contentious. The effect size of treatment differences for bona fide psychotherapies has varied from meta-analysis to meta-analysis; however, when differences are found, they are often in the vicinity of $d = 0.20$ (e.g., Barkham & Lambert, 2021). Such an effect size is conventionally considered "small" (Cohen, 1988); however, such a rule of thumb was only ever meant to be treated as a rough heuristic. Consequently, researchers have contested whether an effect size approximating $d = 0.20$ is

meaningful (e.g., Marcus et al., 2014; Tolin, 2010) or insignificant (e.g., Messer & Wampold, 2002; Wampold et al., 1997).

To specify a relevant effect size, one may consider whether the effect is impactful for individuals—that is, whether somebody participating in psychotherapy is likely to notice a difference subjectively. This threshold is referred to as the "minimally important difference" (MID); a preliminary estimate by Cuijpers, Turner, et al. (2014) suggests a MID of $d = 0.24$ may represent a detectable difference in psychotherapy outcome. Approach to the effect size focused on the impact of treatment on individuals is the common language effect size (CLES; McGraw & Wong, 1992) or "probability of superiority," which represents the likelihood that a person picked at random from Treatment A will have a better outcome than someone picked at random from Treatment B (where 50% is no difference between treatments). A CLES of 55% (i.e, a marginal improvement of 5% over 50-50 chances) is equivalent to $d = 0.18$. An effect size of $d = 0.20$ equates to a CLES of 55.6%.

One may also consider whether the effect is impactful at scale, across individuals in the population receiving treatment. The number-needed-to-treat is one way of expressing this effect, indicating the number of people needed to be treated with Treatment A before one more successful outcome occurs relative to Treatment B. Depending on the formula one uses, an effect size of $d = .20$ is equivalent to an NNT of 9 (Kraemer & Kupfer, 2006) or 13 (assuming the chances of a favourable outcome in the comparator treatment are 50-50; Furukawa & Leucht, 2011). Either of these NNTs compare favourably to NNTs for critical medical interventions, such as the effect of vaccination on preventing influenza in the elderly (NNT of 29; Demicheli et al., 2018), antibiotics for reducing infections in open limb fractures (NNT of 16; Gosselin et al., 2004), blood thinners for reducing stroke (NNT of 25; Aguilar & Hart, 2005); corticosteroids for

reducing death in pneumonia (NNT of 17; Stern et al., 2017); or, more familiarly, ibuprofen for eliminating tension headache (NNT of 14; Derry et al., 2015). To use an example that is arguably more comparable to the case of relative psychotherapy comparisons (e.g., a comparison of two *bona fide* interventions that requiring patient adherence, lifestyle change, between-session work, alliance with the clinician, and motivation), Mediterranean-type diet was compared to the American Heart Association Step 1 diet in heart attack survivors; the study was terminated early due to the relative benefits of the Mediterranean diet being deemed to be substantial: NNT for preventing nonfatal heart attack recurrence was 18, preventing death by any cause was 30, and preventing cancer occurrence was 30 (de Lorgeril et al., 1998). Furthermore, speaking directly to the idea of effects at scale: in 2015, 3% of Canadians aged 12 and over consulted with a psychologist (Canadian Community Health Survey, 2015), representing 1.1 million Canadians. If 50% of these individuals received a "superior" psychotherapy A and 50% an "inferior" psychotherapy B, and if 50% of those receiving B had a favourable outcome (275,333 people), then an NNT of 13 would be equivalent to 318,836 people having a favourable outcome in A – a difference of 43,503 people (Furukawa & Leucht, 2011). These numbers are confessedly somewhat contrived; however, they illustrate a broader point about how a conventionally small effect size can nevertheless be impactful in large-scale health systems.

Yet a further factor may be considered when interpreting the effect size for the question of the DBV. If the DBV is understood as related to treatment families (i.e., the claim that there are equivalent outcomes for CBT vs. dynamic vs. experiential-humanistic therapies vs. ... etc.), then if the DBV were false, one hypothesis could be that regardless of the difference in effects seen when comparing treatments within a family (e.g., one CBT vs. another CBT), the effects between families (e.g., CBT vs. dynamic) should nevertheless be larger. Within-family effects

may be small (e.g., if all treatments from the same family are roughly equally efficacious) or large (e.g., if some treatments within a larger family are more efficacious than others); however, if it is correct to say that the treatment family makes a difference to the outcome, then the relative effect size between families should nevertheless exceed the within-family effect. That is, determining whether the effect size is meaningful when comparing one therapy family to another may not simply be a matter of crossing some size threshold, but it can also be *relative*, anchored to the effect of the effect size of treatments within a family. Such provides additional information in the context of the meta-analysis using absolute effect sizes: by coding within- vs. between-family effects as a meta-regression moderator, information about the impact of therapy families can still be gleaned, despite the fact the meta-analysis otherwise avoids categorizing effects.

For the purposes of this meta-analysis, these various approaches for appraising a relevant effect size will be considered separately. Because any threshold for a "meaningful" effect size will be subjective, these thresholds will be adopted with the intention of serving as merely *reasonable* benchmarks, rather than purporting to being objective or indisputable. To assist in straightforward interpretation of the effect size, both *g* and the CLES will be used throughout the paper.

Following Cuijpers, Turner, et al. (2014), a threshold effect size of $g = 0.24$ will be considered practically meaningful difference in the context of individual patient's qualitative experience—equivalent to a probability of superiority of 56.7% (vs. 43.3% for the alternative). This threshold will be the primary one adopted for the meta-analyses in these studies. In the meta-analyses of absolute values (addressing the question "By approximately how much all psychotherapy treatments equally effective?"), an effect size of $g = 0.24$ or greater would be taken to indicate that, when considering the distribution of true (upper bound) effect sizes for

comparisons between a psychotherapy A and a psychotherapy B in relevant populations, the average of that distribution will be of a magnitude that is itself on average detectable to patients. Similarly, for the standard meta-analysis of various therapy families vs. CBT, an effect size of $g = 0.24$ or greater would indicate that, when considering the distribution of true effect sizes for comparisons of different therapy families in relevant populations, the mean of that distribution will be of a magnitude that is itself detectable on average to individual patients.

As previously identified, the most relevant conceptual issue for therapy families may be whether the effect size between families exceeds that within families (e.g., the effect size for CBT vs. another family exceeds that of one CBT to another CBT). Requiring the between-family effect to exceed the within-family effect means setting a threshold for $g_{between} - g_{within}$. Setting this to g = 0.24 is an option, but a relatively stringent one; if $g_{between} = 0.24$ (a detectable difference between treatments), g_{within} would have to be precisely $g = 0.00$ to meet the threshold. Accordingly, it may be reasonable to slightly liberalize the threshold. Thinking in terms of the CLES, one may deem the differences between therapy families to be reasonably meaningful if the probability of superiority exceeds that of the within-family effect size by a marginal 5%. For example, if CBT were to have an effect size of $g = 0.24$ relative to other families (a CLES of 56.7%), the effect may not be meaningful unless the marginal CLES for CBT-CBT comparisons is at least 5% less, at 51.7% ($g = 0.06$). Adopting this difference as a meaningful one gives $g_{between} - g_{within} = 0.18$ as the threshold by which the between-family effect would have to exceed the within-family effect.

Lastly, a much lower threshold for a meaningful effect size would be reasonable to adopt when considering effects at the population level; indeed, the fact that seemingly minute effects may nevertheless be practically significant across large numbers of people is a well-recognized

phenomenon (e.g., Abelson, 1985; Cohen, 1988). For example, if 1,000,000 individuals are evenly divided between Treatment A or B, and the chances are 50-50 for a favourable outcome in Treatment B, then if Treatment A is superior to B by just $d = 0.05$ (NNT of 50) this would equate to 10,000 additional people with favourable outcomes (Kraemer & Kupfer, 2011). Such an effect size is arguably a reasonable standard for being practically meaningful at scale. However, it is also the case that every analysis conducted in this paper is unlikely to be sufficiently powered to detect such an effect. For example, assuming large heterogeneity, the smallest effect that could be detected by the most expansive meta-analysis conducted in this paper (at $\alpha = .05$, $1\text{-}\beta = .80$) is just $g = 0.05$. Accordingly, we may operate from the position that, if one of the meta-analyses in this paper is sufficiently powered to detect a statistically significant effect, then that effect will clear the threshold for practical significance across large numbers of people. Because this standard is arguably a liberal one, it will be considered as secondary to the MID-based threshold of $g = 0.24$.

Aims of the Present Study

With this conception of the DBV, of psychotherapy, and of relevant effect sizes in mind, the goals of this meta-analytic paper may be specified. The primary aim of this paper was to

1. Determine, as a global test, whether there are generically differences in relative efficacy between *bona fide* psychotherapies (without classifying into families); and
2. Estimate the differences in efficacy between *bona fide* psychotherapy families (i.e., classifying treatments into families and using one therapy family as a reference class), and appraise whether the differences between therapy families meaningfully exceed that of the differences within a family.

Additionally, a secondary aim was to

3. Assess the extent to which the meta-analytic results are sensitive to various conceptions of the DBV and corresponding methodological choices (e.g., magnitude of the between-treatment effect of unclassified treatments, via meta-analysis of absolute values, effect size of treatment families relative to one another via directional effect sizes; e.g., Marcus et al., 2014; Tolin 2010).

To answer these questions, this meta-analytic study was undertaken in three parts, adopting three different methodologies. In the first part, the primary objective was to address Aim 1. The methods of Wampold et al. (1997) were applied to determine whether, in general, there was statistically significant variation between the effects of different treatments when compared head-to-head. That is, the effect sizes for each of the comparisons between two treatments were randomly assigned either a positive or negative sign, producing a distribution of effect sizes centred around zero, and the degree of homogeneity of the effects was assessed to appraise whether treatment differences were clustered closely around zero (indicating minimal between-treatment differences) or widely dispersed (indicating large treatment differences).

The second part primarily addressed Aim 3: the magnitudes of the differences between treatments were estimated using the absolute values of effect sizes, as an analogue to the standard meta-analysis but for unsigned and unclassified effect sizes as used in part one. Meta-regression techniques were used to determine whether variance in absolute effect size could be explained using four key moderating variables: whether comparisons were within a therapy family (e.g., CBT vs. CBT) or between families (e.g., CBT vs. IPT); the DSM-5 disorder population from which the study's sample was drawn (e.g., depressive disorders, anxiety disorders, etc.); the difference in researcher allegiance score between comparators; and study risk of bias per the Revised Cochrane Risk-of-Bias tool (RoB2; Sterne et al., 2019). Because the

absolute effect sizes may overestimate the actual effect size magnitude (Marcus et al., 2014; Wampold et al., 1997), these meta-analyses are intended as supplemental to the meta-analyses of directional effect sizes (below).

The final part was intended primarily at addressing Aim 1, applying the methods of the "cognitive contrast" of Marcus et al. (2014): the most frequent comparator (CBT) was used as a control condition and compared against other therapy families to estimate ordinary (i.e., non-absolute valued) effect sizes. Again, meta-regression techniques were used to assess the impact of disorder, researcher allegiance, and RoB2 score. The impact of the comparator to CBT (i.e., IPT, PDT, another CBT, etc.) was used as the fourth moderator in this analysis, rather than the impact of within- vs. between therapy family comparisons.

Method

Location of Primary Studies

Although comprehensive database searches are standard in almost all contemporary meta-analyses, they are best suited for inquiries into the DBV that are limited to a specific disorder. In contrast, for tests of the DBV that aim to consider *all* psychotherapy comparisons (as in this study), obvious search terms such as "comparative" are frequently not found in the title, abstract, or indexed subject-terms of relevant trials, and other obvious search terms such as "therapy" do not adequately distinguish between relevant and irrelevant studies (Swift & Wampold, 2018). Consequently, relevant studies were located using the three methods outlined below. To minimize redundancies, the title/abstracts of primary studies collected with each of these three methods were reviewed to eliminate duplicate studies. Afterwards, screening of the retained studies proceeded as described in the following paragraphs.

The first search method involved a manual search of several key journals between 1980 and 2020[5], inclusive: *Behavior Therapy*, *Behaviour Research and Therapy*, *Journal of Consulting and Clinical Psychology*, *Journal of Counseling Psychology*, *JAMA Psychiatry*, and *Cognitive Therapy and Research* (journals selected per the methodology of previous meta-analyses, e.g. Marcus et al., 2014; Wampold et al., 1997), as well as *Psychotherapy* and *Psychotherapy Research* (selected to supplement the above list with journals that focus less heavily on CBT therapies). The year 1980 was used as a starting point due to this roughly representing the advent of what has been described as Generation III psychotherapy research—that is, clinical trials for psychotherapy research, in which randomized controlled trials of psychotherapy began to be undertaken in specific clinical populations (Goldfried & Wolfe, 1996; Wampold et al., 1997). Generation III research (and beyond) is of particular use for questions of relative outcome due to the emphasis on reliable diagnostic categories and the use of *bona fide* clinicians and clinical populations (rather than university student clinicians and participants). These articles were screened based on whether their abstracts or titles give an indication that they compared psychotherapy outcomes; to be retained for possible inclusion, there must have been indication that at least some of these were direct, head-to-head comparisons. Moreover, only English-language studies of individual adult therapy were retained. Following this screening, articles were examined in detail and included only if they met the following criteria: the studies must (a) be randomized controlled trials, (b) involve direct comparisons of (c) "*bona fide"* psychotherapies, (d) evaluate therapies in populations with diagnosed psychological disorders, and (e) have no apparent confounds with medication use across the therapies being evaluated.

[5] This search stop date was due to practical considerations rather than scientific grounds—e.g., disruption of university access and normal routine due to COVID-19 pandemic, volunteer RA stepping down, retirement of supervisor (2021), clinical internship (2022), health-related leave, etc.

Studies were excluded if they (a) only provided a re-analysis of a dataset that has already been included (but any analysis of new outcome measures was retained) and (b) did not include information from which effect sizes can be calculated (e.g., means, sample sizes, variances). These criteria are discussed in further detail below. Covidence, a software partnered with Cochrane, was used to collect the abstracts and to assist with the title/abstract and full-text screenings.

Second, to capture unpublished data in the grey literature, a search of dissertations was undertaken using a systematic term search of the ProQuest Dissertations and Theses Global database (see Appendix B for the terms used in this search). Abstracts collected based on this search algorithm search were then screened, as before, for indication that psychotherapy outcomes were directly compared at some part of the study. Following this title/abstract screening, the full-text studies were screened and retained only if they meet the criteria described above. As before, Covidence was used to aid with study collection and screening.

Third, relevant studies were located using past meta-analyses on the subject. A systematic literature search was undertaken to comprehensively locate meta-analyses published between 2006 and 2021 relevant to evaluating the DBV. This was accomplished through a search of indexed subject-terms and title/abstracts in the PSYCInfo database (see Appendix B for the terms used in the search algorithm). During the screening phase, these meta-analytic studies were excluded if their abstracts or titles did not give clear indication that they investigated primary studies of psychotherapy outcome and if they did not give clear indication that at least some of these investigations were direct, head-to-head treatment comparisons. Following the screening, the studies recorded in the reference list of each of these meta-analyses were then be reviewed and included if they meet criteria described above.

A randomly selected 10% of studies was scored by an undergraduate volunteer at the title/abstract search phase, and by the research supervisor at the detailed inclusion/exclusion phase. For the title/abstract search, interrater agreement was 88%, with a Cohen's kappa of 0.75. For the detailed inclusion/exclusion phase, interrater agreement was 89%, with a Cohen's kappa of 0.64. Disagreement was resolved through discussion; consensus was reached in nearly all instances, and the author served as the final arbiter of inclusion eligibility in instances when it was not.

Inclusion Criteria

In line with previous meta-analyses on the DBV (e.g., Tolin, 2010; Wampold et al., 1997), studies were limited to English language studies for practical reasons. Studies were also limited to those of individual therapy for adults. Although the DBV controversy extends to interventions for children and adolescents, and to group interventions as well as individual ones, the scope of these meta-analyses were limited to minimize "apples-and-oranges" comparisons (e.g., Eysenck, 1994).

Primary studies were limited to randomized controlled trials to ensure that conclusions drawn on the basis of these meta-analyses were as free from confounding variables as possible—that is, to ensure that the aggregated, summary effects of these analyses represent the clearest possible indication of the relative effects of the interventions, putting aside all other considerations.

Direct Comparisons

Studies were limited to direct comparisons of psychotherapy (as opposed to indirect comparisons of therapies vis-à-vis a comparator condition, such as treatment-as-usual (TAU) or a wait-list control). In general, direct comparisons have two major benefits over indirect

comparisons. First, they avoid the problem of non-equivalent control groups. In many indirect studies of psychotherapy effects, TAU is understood to function as a control condition, purportedly allowing the authors to demonstrate whether the psychotherapy in question is better than treatment "as usual." However, the norms for care "as usual" vary substantially from study to study, from treatment centre to treatment centre, and over time (e.g., Wampold et al., 2011), meaning that TAUs cannot be considered equivalent comparator conditions from one study to another. Directly comparing psychotherapies to one another obviates this issue. Second, direct comparisons also remove the need for statistical controls of various confounds that might otherwise present a threat to the validity of the comparative data (e.g., confounds by setting, by treatment implementation, or by measurement reactivity; Wampold et al., 1997). In direct comparisons, both therapies under consideration are conducted in the same setting(s), with the same measurements and, theoretically, the same quality of implementation (but *pace* quality issues related to allegiance bias; Munder et al., 2011); this is not the case for indirect comparisons, where each comparator therapy may be undertaken in different settings from their respective controls. It is worth noting that indirect comparisons relative to a wait-list control—unlike those relative to TAU—are not plagued by issues of control non-equivalence, as the experience of waiting for treatment is unlikely to vary systematically from setting to setting. However, other confounds (such as different referral sources or recruitment strategies across settings) may still present a problem, again making direct comparisons an arguably more optimal choice for a DBV comparison.

Bona Fide Criterion

Limiting studies to those of "*bona fide* psychotherapies" was a strategy first devised by Wampold et al. (1997) and has since become a benchmark for any meta-analysis of the DBV.

These criteria are necessary to ensure that psychotherapies being compared are true, real-world "competitors," rather than treatments that are intended to fail so as to prove the viability of another treatment (Wampold et al., 1997) or are composed of only a few elements that are likely to be therapeutic. This analysis followed the approach of Marcus et al. (2014), who loosened these criteria to allow studies in which therapy was delivered by trainees without a master's degree, but under the supervision of a doctoral-level mental health professional (e.g., a licensed psychologist or psychiatrist). This liberalizing of the criteria is consistent with evidence that therapist experience is uncorrelated with client outcome in general (e.g., Sylvan et al., 2022), and that student therapists have no worse outcomes than do qualified therapists (e.g., Goldstein et al., 2020; Mason et al., 2015).

Although less strict with the licensure status of the therapist, the current analysis was stricter with another of Wampold et al.'s (1997) criteria—rather than allow studies of treatments for all types of clinical problems, the current analysis limited studies to those examining psychotherapy for clinical mental disorders per the criteria of the *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed. or later; *DSM–III*; American Psychiatric Association [APA], 1980) or of the *International Classification of Diseases* (9th ed. or later; *ICD*-9; World Health Organization, 1979), as determined by psychologists or psychiatrists (e.g., by chart review; by diagnostic interview). This is not to suggest that nondiagnosable problems are not an important focus for research or clinical practice; however, their inclusion in this study could constitute a potential source of variability in the meta-analytic data and so excluding them is a straightforward way of improving internal validity.

In light of these modifications, to be considered *bona fide* for these meta-analyses, treatments:

1. Must be offered by genuine psychotherapists (i.e., professionals with at least a master's degree—or graduate students under the supervision of psychologists or psychiatrists—who make active treatment decisions);
2. Must be designed for helping with a (broadly construed) clinical problem (as opposed to, for example, being intended solely for use as a control condition in a study);
3. Must be evaluated in a clinical population, i.e., one with individuals with diagnosed psychological disorders (criteria from DSM-III or later, or ICD-9 or later);
4. Must be used in the study to (a) treat features of the aforementioned diagnosed disorders or (b) produce more general improvement or growth in psychosocial adjustment or well-being. Therapy cannot be used exclusively to treat a problem with no close relationship to the diagnosis (e.g., cannot only be geared toward treating obesity in people with major depressive disorder).
5. Must meet at least two of the following criteria: the studies should
    a. cite an established psychotherapeutic approach; or
    b. cite established psychological principles in its rationale; or
    c. be manualized, or codified in a professional book; or
    d. have their active therapeutic ingredients identified and cited.

Additionally, when interventions were classified into therapy families, studies were excluded as non-*bona fide* if they did not meet the minimum threshold criteria for belonging to at least one family of therapies (see "Moderators").

The PRISMA flowchart describing the inclusion process and the reasons for exclusion during full-text screening is presented in Appendix C. A total of 29,435 records were examined (27,798 after removing duplicates) and 1,543 were retrieved for full-text screening. 1,341 were

excluded, leaving a total 187 studies. Initially, some studies included unique outcomes but from the same participant pool as other studies; additionally, there were often multiple effects per study (a total of 2,719 effect sizes). After synthesizing effect sizes for each meta-analysis to account for these dependencies, 165 independent effects were ultimately available, with different analyses based on different clusters of these effects as appropriate; the largest subset of effects, used in the meta-analysis of randomly assigned effects, was based on 158 effects.

The Risk of Bias 2 scores for each primary study, including the overall score, is presented in Appendix D.

Primary Analysis

Coding Criteria

Outcomes and Timepoints. A study's effect was coded as belonging to a primary symptom measure if the measure was used to evaluate symptoms related to the disorder category that characterizes to the study's target population (i.e., measures of depressive symptoms in major depressive disorder). A secondary symptom measure was coded as such when the measure evaluated psychological symptoms other than the primary diagnosis (e.g., of comorbid disorders, conceptually related disorders, etc.). A global measure of functioning was coded as such when it evaluated non-symptom outcomes (e.g., general distress, quality of life, self-esteem, functional impairment, social functioning). Within a given category of outcome variables, therapeutically desirable outcomes were coded with positive values; consequently, measures where higher values represent less favourable outcomes (e.g., a measure of general distress where higher scores represent greater distress) had the signs of their values reversed. Separate meta-analyses were undertaken for each of these three categories.

Measures were also coded as occurring at post-treatment or follow-up. Whereas post-treatment was defined as measures taken immediately at termination, follow-up measures were defined as any measure taken six months after termination or later. Effects from post-treatment and follow-up time points were meta-analyzed separately from each other. Measures taken between zero and six months after termination were not included in the analysis. This is because, if treated as post-treatment data, they would likely be confounded by maturation or history in a way that data collected immediately post-treatment would not be; however, if they were treated as follow-up data, insufficient time is likely to have elapsed to test for the robust perdurance of the effect. If a study included multiple follow-up time points (e.g., six months post-therapy, one year, two years), they were all coded as follow-up but as separate effects; they were later combined into a composite follow-up effect.

If multiple outcomes were reported within a study from within same meta-analysis (e.g., two primary symptom measures at post-treatment), a combined effect for these measures was meta-analyzed, which was synthesized using the mean effect and a pooled variance estimate. A correlation between the measures of .70 was assumed for primary symptom measures (a number based on the correlation between the Beck Depression Inventory – II [BDI-II] and the Hamilton Depression Rating Scale–Revised [HAM-D]; Beck et al., 1996). The same correlation was assumed for multiple secondary symptom measures; because higher correlations between measures are related to more conservative estimations of effects, keeping the correlation as it would be for two measures from the same disorder is likely to have reduced the chances of overestimating the effect (relative to adopting the correlation between less closely related measures). For global measures, a correlation of .60 was assumed, based on the correlation between the Social Adjustment Scale (SAS) and the Outcome Questionnaire–45 (OQ-45;

Lambert et al., 1996). For the overall analysis where measures were combined, an intermediate correlation of .65 was adopted.

When higher scores on a measure indicated less pathology, increased adaptive functioning, or improvement on a favourable outcome, the measure was assigned a valence of +1. When higher scores on a measure represented increased pathology, decreased adaptive functioning, or deterioration on a favourable outcome, it was assigned a valence of -1. As a result, for the meta-analyses where the valence was not randomly re-assigned (e.g., the cognitive contrast meta-analysis), larger summary effects indicated favourable outcome.

When outcome measures used in a study were not conceptually clear or minimally traceable, they were not included in the meta-analysis. Specifically, if a study created an idiosyncratic and untraceable composite measure made from items from multiple different questionnaires, without listing the included items or providing a citation for their procedure, they were excluded. Measures of therapy process (i.e., measures of intermediate factors that purport to drive change in psychotherapy, but which are not themselves outcomes of client improvement outcomes) were also not included as outcomes. Examples include measures of mindfulness, insight, or the presence of cognitive distortions. When outcomes could be construed as both a therapy process measure and a client outcome (e.g., self-esteem, perseverative negative thinking), the outcome was included and coded as a primary, secondary, or global measure, as appropriate.

Effects were initially coded separately as intent-to-treat (i.e., an effect derived from all randomized participants; e.g., last observation carried forward for drop-out participants) or completer (i.e., an effect derived only from those participants who completed the study; e.g., drop-out participants excluded), with an eye to evaluating these as separate subgroups. However,

to prevent redundancy with the Risk of Bias moderator, which considered whether a study uses ITT or completer methods, this distinction was dropped when calculating the summary effect for the meta-analysis. When a single study reported results from both completer and ITT samples, these data were not combined into a synthetic effect to avoid problems with "double counting" results that are closely correlated. In such cases, the ITT comparisons were preferred, and completer analyses were excluded from analysis; the result is a likely more conservative estimate of treatment differences.

Moderators. Included studies were coded for the following possible moderating variables: disorder; methodological quality; therapy comparison-type; and researcher allegiance. These moderators were selected because of their prominence in the psychotherapy efficacy literature. The role of researcher allegiance to a particular therapy orientation has long been recognized as a factor in the psychotherapy outcome literature (e.g., Luborsky et al., 1999; Smith et al., 1980), and although debate has ensued surrounding causal directionality (i.e., whether better outcomes lead to stronger allegiance or vice-versa), there is some evidence to imply the causal primacy of allegiance (Munder et al., 2012). Similarly, concerns about methodological quality affecting the results of meta-analyses (i.e., poorer quality studies showing different patterns of results from high-quality studies) has been a concern since the outset of meta-analytic outcome equivalence testing (e.g., Eysenck, 1978; Wilson & Rachman, 1983). As previously described in the introduction, even contemporary meta-analyses may fail to account for these factors. Furthermore, undertaking meta-regression on both researcher allegiance *and* methodological quality is critical, given previous findings that the effect of researcher allegiance is highest when methodological quality is lowest (Munder et al., 2011). Also as previously outlined, certain influential earlier meta-analyses (e.g., Wampold et al., 1997) have not

accounted for the heterogeneity between studies (a) that compare treatments within the same therapy family versus those comparing treatments between families (e.g., CBT-CBT comparisons vs. CBT-interpersonal therapy comparisons) or (b) due to the clinical population in which they were assessed (e.g., depression vs. eating disorders). Consequently, in keeping with more contemporary meta-analyses on the topic (e.g., Marcus et al., 2014), it was important to account for these key factors.

Disorder was coded based on which clinical populations are cited in the population inclusion criteria by each study's authors in terms of a DSM or ICD diagnosis (studies that did not include such a citation were excluded earlier by the screening criteria). Because disorder type was a categorical variable, it was dummy-coded for use as a moderator in the meta-regression, with one category being used as a reference treatment family (usually "Depressive Disorders"). For the sake of statistical power and reducing the quantity of dummy-coded variables, the individual disorders were coded into broader categories (e.g. "Anxiety Disorders," "Trauma- and Stressor-Related Disorders") based on a DSM (5th ed.; *DSM–5*; APA, 2013) conceptualization of each disorder class. If a study used more than one diagnosis for its population inclusion criteria, and the comparisons for each diagnosis were separated out (e.g., provides BDI-II outcome for depressed patients and a separate BDI-II outcome for anxious patients), then each outcome was coded as a separate effect associated with the relevant diagnosis. If a study used more than one diagnosis for its population inclusion criteria and outcomes were not separated by diagnosis, then the diagnosis that was most frequent in the sample was coded (e.g., if 52% of the sample had MDD and 48% had GAD, MDD would be coded). Comorbid diagnoses were not considered in the coding unless the comorbidities defined the clinical population per the authors' inclusion criteria.

Following contemporary meta-analyses of psychotherapy outcome (e.g., Cuijpers, Miguel, Ciharova, Harrer, et al., 2024; Papola et al., 2024), sources of methodological bias were assessed using the Risk of Bias 2 tool (Stearne et al., 2019). This tool appraises sources of methodological bias including those arising from randomization, assignment to or adherence to the intervention, missing outcome data, measurement of the outcome, and selection of the reported result. For this study, the measure was adapted by removing some criteria related to blinding, and some criteria related to measurement bias, as they are inappropriate to psychotherapy trials (e.g., therapists providing intervention cannot be blind to treatment). Additionally, when assessing bias in the per protocol effect (i.e., adherence to treatment), studies were assessed for the occurrence of non-protocol interventions and failures in implementing the intervention that could have affected the outcome, but not non-adherence to their assigned intervention by trial participants. See Appendix E for full criteria.

In order to determine therapy comparison-type, each treatment cited by each study was first categorized into key families: Cognitive-Behavioural Therapies (CBT), Psychodynamic Therapies (PDT), Interpersonal Therapy (IPT; e.g., Weissman et al., 2000), and Experiential-Humanistic Therapies (EHT), and Integrative therapies. Eye Movement Desensitization and Reprocessing (EMDR; e.g., Shapiro, 2001) was added later. Rather than follow Tolin (2010), who identified therapy families based on personal expertise, Baardseth et al. (2013), who employed an empirical method, or Marcus et al. (2014), who relied on each study's authors, the families for this meta-analysis were coded based on criteria associated with the actual techniques associated with each therapy. Such a classification technique gives insight into the questions most central to the DBV (e.g., the relations between therapeutic approaches, techniques, and

outcomes) and helps to circumvent potential hurdles associated with "mislabelling" (e.g., a treatment with minimal CBT techniques being labelled by its authors as a CBT).

To this end, the Multitheoretical List of Therapeutic Interventions (MULTI; McCarthy & Barber, 2009) was used to help define the various therapy families. The MULTI is a 60-item measure used by therapists, clients, or observers to rate therapy interventions, scoring the techniques on subscales corresponding to seven therapeutic orientations (behavioral, cognitive, dialectical-behavioral, interpersonal, person-centered, psychodynamic, process-experiential) as well as common factors techniques. The MULTI has several properties that made it useful for the following study: first, it provides subscales scores that correctly classify interventions to their purported therapeutic orientations with a 0% error rate for therapy-experienced raters and 11%-12% for therapy-naïve raters (McCarthy & Barber, 2009); second, although it was developed for clients or therapists to rate their own sessions, or for observers to rate videotapes, the MULTI focuses on therapist actions (rather than their intentions or on the theoretical rationale)—making it straightforward to adapt to rating the descriptions of therapy techniques provided in manuscripts; third, it distinguishes between (a) common factor processes and those specifically associated with humanistic techniques, and (b) between those specifically associated with IPT and with interpersonal-psychodynamic therapies—distinctions not always drawn; and lastly, it uses jargon-free language, making it appropriate for use by untrained and therapy-naïve research assistants. For this study, the observer form of the MULTI was adapted in the following ways: (a) it was used to rate descriptions of interventions provided in manuscripts rather than videotapes; (b) the common factors subscale was not scored, as *ex hypothesi* these scores would be common, and hence orthogonal, to therapeutic orientation; (c) the cognitive, behavioural, and DBT subscales were collapsed to make a higher-order "CBT" subscale; (d) the person-centred

and process-experiential subscales were collapsed to make a higher-order "Experiential and Humanistic" subscale; and (e) the scores—originally, from 1 to 4, corresponding to "not at all typical of the session" to "very typical of the session," were modified as follows. A score of 1 indicated "Not a match" – that the intervention doesn't fit the item, or there was no information to suggest a match. A score of 4 indicated that the intervention matched the item. A score of 2 or 3 were available if the intervention description was a weak partial or strong partial match to a MULTI item, respectively; however, raters were encouraged to score "1" or "4" whenever possible. See Appendix F for the adapted MULTI.

To be coded as a match to CBT, EHT, IPT, or PDT, the intervention under consideration had to score a minimum average score of "2" on the adapted MULTI subscale (i.e., at least a weak partial match on average), and had to be at least 0.5 points higher on average than any other subscale. When an intervention had an average MULTI score above 2 for two or more subscales, and none of these scores exceeded any other by 0.5 points or more, it was classified as an "Integrative" therapy. If it scored below 2 for all categories, it was excluded as a "non-bona fide" therapy (see "Inclusion Criteria").

The cognitive, behavioural, and DBT subscales were incorporated together due to the shared theoretical background and longstanding historical connection between these treatments (e.g., as "waves" of a shared tradition; Hayes & Hofmann, 2017; Thoma et al., 2015). The view that these specific therapies are part of a larger shared umbrella is reflected in both the attitudes of the proponents of these therapies (e.g., Beck, 2007; Hayes, Follette, & Linehan, 2004; Linehan, 1993) and in professional associations that group them accordingly in order to represent them (e.g., the Association for Behavioral and Cognitive Therapies).

When coding the studies, it became clear that a non-negligible minority of interventions were classified by study authors as "EMDR." These studies were uniformly coded as "CBT" by the MULTI, consistent with some contentions about the active mechanisms of EMDR (e.g., Cuijpers et al., 2020; Lohr et al., 2015; Rosenfeld & McLean, 2023). However, in contrast with the interventions represented by the cognitive, behavioural, and DBT subscales, the theoretical rationale for EMDR (i.e., the "adaptive information processing" model; Shapiro 2001) is not held in common as part of the shared CBT historical-theoretical tradition. Moreover, the purported active technique (viz. bilateral stimulation) is not a standard CBT technique nor is it captured by the MULTI. Consequently, a separate category for EMDR was created. Interventions were coded as EMDR if they cited the EMDR manual by Shapiro without modification, if they cited bilateral stimulation as a core intervention, or if Adaptive Information Processing theory was cited as the primary rationale for the treatment.

Having categorized each treatment examined in each study, every effect size was then coded as either a within-therapy family comparison (one where two treatments are from the same study; e.g., a CBT-CBT comparison) or a between-family comparison (e.g., a CBT-PDT comparison). This was necessary for the meta-analysis of absolute value effect sizes, because every study yields effect sizes that concern the comparison of *two* therapies—it would therefore be inappropriate to code an effect size as "belonging" to any particular treatment family. This rating of comparison-type (i.e., whether the effect is based on a within- versus a between-family comparison) was then used as a covariate in the meta-regression for the meta-analysis of absolute valued effect sizes.

Lastly, researcher allegiance to each of the therapy families (at the time of each study) was coded according to a scheme introduced in Yulish et al.'s (2017) meta-analysis. A score was

assigned for each treatment in a study, ranging from +3 (representing allegiance in favour of the treatment) to -3 (representing allegiance against the treatment). The final allegiance score for each comparison of Treatment A vs. Treatment B was calculated by subtracting the allegiance score for Treatment B from that of Treatment A. One item from Yulish et al.'s original system was removed (viz., "+1 if greater face-to-face dosage compared with other treatment") as difference in face-to-face dose was coded separately as its own potential moderator. See Appendix G for the items included in the coding system. For this study, allegiance was coded on the basis of the principal investigator (and not, say, of the therapists who provided services in each study), as previous findings (e.g. Tolin, 2010) have suggested that only the lead investigator's allegiance has a relation with study outcome.

In addition to the planned moderators, two post-hoc moderators were also coded: year of publication of the study, and difference in dose between treatments, calculated in hours. Ultimately these were dropped due to considerations of statistical power, but reintroduced in one analysis (i.e., the third meta-analysis [cognitive contrast], for primary disorder measures at termination) to replace the methodological quality moderator (see "Cognitive Contrast, Meta-regression and subgroup analyses, Termination, Primary disorder-specific outcomes").

Plan for Statistical Analysis

Meta-analysis was used to calculate a weighted average of the differences between treatments in the included primary studies. Individual effects were weighted by precision (i.e., inverse of their variance). Treatment differences are expressed as a standardized mean difference, with a correction to prevent overestimating the absolute value of the effect in small samples (i.e., Hedges' *g*; Hedges, 1981), along with 95% confidence intervals (95% CIs). When study effect sizes were given as other effect sizes (e.g., odds ratios), they were converted to

Hedges' *g*. The random-effects model was employed for all analyses: that is, the studies in each analysis are assumed to be a random sample from the space of all possible studies with comparable populations, and this analysis makes inferences to that space (Borenstein, 2019; Borenstein et al., 2010; Borenstein et al., 2021; Hedges & Vevea, 1998; Higgins & Thomas, 2019). The summary effect size and 95% CI are thus modelled as the estimated mean effect size in the space of all comparable studies, rather than simply representing the mean of the included studies. The *Z* test, which was used to assess the statistical significance of the summary effect size, tests the null hypothesis that in populations comparable to those in the analysis, the mean effect size is zero.

The dispersion of the data was tested using Cochran's *Q* statistic, which assesses the null hypothesis that all studies in the analysis share the same true effect size; a significant test indicates that the summary effect reflects multiple, or a continuum of, true effect sizes. The degree of spread is reflected in the estimate of the τ^2 and τ statistics, which indicate the variance and standard deviation of true effect sizes, respectively. Assuming that the true effects are normally distributed in units of *g,* approximately 68% of true effect sizes fall between $\pm\tau$ of the mean effect, and approximately 95% of true effect sizes fall between $\pm 2\tau$ of the mean. This latter interval, in which the true effect size falls in 95% of all comparable populations, is referred to as the prediction interval. The I^2 statistic was calculated to determine what proportion of the observed variance reflects true variance in effect size (as opposed to sampling error)—that is, the ratio of τ^2 to total variance. The greater the I^2, the greater the relative inconsistency among the included studies.

To account for dependencies in the data (i.e., multiple comparisons from the same sample), synthetic effect sizes were produced using the mean effect of the multiple comparisons

and a pooled variance estimate with a plausible estimate of the correlation between the measures. To wit, a correlation between the measures of .70 was adopted for primary and secondary symptom measures, a correlation of .60 was adopted for global measures, and an intermediate correlation of .65 was adopted when measures were combined (see "Coding criteria" above for rationale and details). When studies reported results from both completer and ITT samples, ITT comparisons were preferred and completer analyses were excluded (see "Coding criteria" above for rationale and details).

Outliers in meta-analysis have often been identified by locating studies in which the 95% CI of their effect size does not overlap with the 95% CI of the summary effect size (e.g., Cuijpers et al., 2014). However, under the random effects model, studies are hypothesized to come from a distribution of true effects with variation, and the 95% CI of the summary effect represents only the accuracy of the estimate of the mean—not the extent of the variation of true effect sizes. Non-overlap with the 95% CI of the summary effect—especially when error is small—may therefore be expected under the random effects model. Consequently, studies were instead excluded when the 95% CI of their effect size did not overlap with the 95% prediction interval (PI)—that is, the estimate of the variation in the distribution of true effect sizes itself. Deletion (rather than Windsorizing) was deemed appropriate for certain outlying effects that appeared aberrant to the point of being methodologically suspect (e.g., Farahimanesh et al., 2021, had an effect size of $g = 6.80$ in the meta-analysis of absolute effect sizes); for consistency, deletion was then applied for all outlying effects.

When warranted by tests of heterogeneity (i.e., significant Q test; substantial values of τ and I^2 [i.e., moderate or greater levels per the criteria of Hedges & Piggott, 2001, and Higgins et al., 2003]), meta-regression was used to assess the extent to which the spread of the data could be

explained by moderating variables. Meta-regression is the extension of regression techniques to the meta-analytic context, estimating the extent to which a variable predicts effect size when holding other variables in the model constant. For continuous variables, this is expressed as the change in effect size (in *g* units) associated with a unit increase in the variable (e.g., "there was an increase of 0.2 *g* for every unit increase in researcher allegiance score"). For categorical variables, this is expressed as the difference in *g* units relative to a comparator category (e.g., "the average in the anxiety disorder population was 0.2 *g* higher than the average in the depressive disorder population"). The *F* statistic was used to test the overall model (i.e., the null hypothesis that none of the covariates explain any variation in effect size). The *Q* statistic was used to test the goodness-of-fit (i.e., the null hypothesis that unexplained variance is zero) and τ^2 and I^2 were calculated as absolute and relative measures of the true variance, respectively. An analogue of R^2 was calculated as the ratio of the true variance explained by the model to the total amount of true variance. To assess the significance of each moderator, the *t* and *F* statistics were used to test the null hypothesis that a given covariate was not independently related to effect size (i.e., that it is unrelated, or that its effect is explainable as a confound of another variable in the model). If the meta-regression suggested that a categorical variable was independently related to effect size, a subgroup analysis was undertaken to estimate the effect size in each separate group (e.g., the effect size in each therapy family, or in each disorder population).

Statistical power for the main effect (*Z* tests) was based on the normal distribution and calculated using the estimated mean effect size and the square root of its variance, with a criterion $\alpha = .05$. Power for the heterogeneity (*Q*) test was based on the χ distribution and calculated using degrees of freedom and the non-centrality parameter. For the *Q* test, the criterion $\alpha = .10$ was used, which is conventional due to low sensitivity of the test (Fleiss, 1981).

A maximum type II error rate of $\beta = .20$ was accepted for the analyses, both due to prevailing convention and to adopt a conservative approach to rejecting the DBV. When an analysis did not achieve a power of $1- \beta = .80$, the significance test was not reported. However, the parameter estimates (i.e., effect size, confidence interval, prediction interval, and τ) were reported. The meta-regression was considered powered at the minimally sufficient level when, per the Cochrane Handbook, the meta-analysis included 10 studies per moderator (Higgins & Green, 2011, 9.6.5.1); an attempt was also made to ensure that each covariate had a minimum of ten included effect sizes. When this minimum threshold was not met, categories were collapsed to reduce the number of covariates in the model, or the analysis was not conducted.

All computations were carried out using Comprehensive Meta-Analysis Version 4 (Borenstein et. al., 2022).

Additional Analyses

Publication Bias

One concern when evaluating meta-analytic results is the possibility that studies with small sample sizes may have systematically larger effect sizes than studies with larger sample sizes (i.e., the "small-study effect"). This may be indicative of publication bias: small studies with small effect sizes may not be published, artificially leading to a preponderance of small studies in the literature with large effect sizes. Addressing the small-study effect assists in making determinations as to whether the effects uncovered by the analysis can confidently be ascribed to treatment differences, as opposed to having drawn upon a skewed initial pool of studies. This may be especially important if the allegiance effect holds true, as bias towards one particular therapy family may prevent authors from publishing studies which seem to "count against" their chosen therapy family (i.e., the "file-drawer effect").

One method to reduce such bias, employed in this study, is the inclusion of high-quality "grey literature" in the form of unpublished dissertations. Additionally, the small-study effect may be assessed through statistical means. Duval and Tweedie's (2000) trim-and-fill procedure searches for asymmetry in the plot of effect size vs. standard error, as would be seen if there are a high number of small studies with large effect sizes and an absence of small studies with smaller effect sizes. It then "trims" the asymmetric studies to locate an unbiased estimate of the effect, and then re-inserts them while simultaneously imputing their inverted values. In this manner, the procedure generates an effect size estimate that "corrects" for the hypothesized missing studies. Additionally, Egger et al.'s (1997) test of the intercept uses regression techniques, predicting the standardized effect of bias from the inverse of the standard error. These methods were used for the third part of this meta-analytic study (the replication of the "cognitive contrast" comparing CBT against other therapy families, using ordinary, non-absolute valued effect sizes).

For the second part of the meta-analytic study (i.e., determining the magnitude of differences between treatments using the absolute values of effect sizes), the distribution of effect sizes was fundamentally biased by design (i.e., no effect size below zero was included). This built-in bias confounds the usual tests of publication bias such as Egger's test and Duval and Tweedie's trim-and-fill procedure. Consequently, a cumulative meta-analysis was conducted to assess for the small-study effect. In cumulative meta-analysis, average effect sizes are calculated by starting with the largest, most precise study and successively adding smaller and less precise studies. By comparing the average effect when including only the 50% of the studies that are the largest with the average effect when also including the smaller, less precises studies, the impact of these smaller and potentially biased studies on the overall effect can be determined.

Results

Search and Study Characteristics

A total of 165 independent RCTs were included, representing a total of 13,370 participants. The number of participants per study ranged from 12 to 440 (median: 65). The demographic information of the participants in the primary RCTs is presented in Appendix H. In summary: the average age of participants ranged from 19 to 79 (median: 37) years. Forty-six per cent of studies were conducted in North America, 46% in Europe, and 8% elsewhere. Forty per cent of studies had more than half ethnically White participants, including 10% of studies that had a sample with 90% or more White participants; 51% reported no data on the ethnicity of the participants. Thirty per cent had between 40-60% men participants; 56% had fewer than 40% men participants; and 12% had greater than 60% men participants (3% did not report the gender of their participants). Seven per cent of studies examined participants with addictive disorders; 33% examined participants with anxiety disorders; 20% depressive disorders; 5% eating disorders; 2% neurodevelopmental disorders; 8% obsessive-compulsive/related disorders; 4% personality disorders; 1% sleep disorders; 1% somatic disorders; and 18% trauma disorders. Sixty-five per cent of comparisons were between treatments within the same therapy family (e.g., CBT vs. CBT, PDT vs. PDT), whereas 35% compared treatments between different therapy families (e.g., CBT vs. PDT). Of the comparisons in which a CBT treatment was the first comparator, 66% had another CBT treatment as the second comparator; 8% had EHT; 3% had EMDR; 9% had an Integrative treatment; 7% had IPT; and 8% had PDT.

Overall Differences (Randomly Assigned Effect Directions)

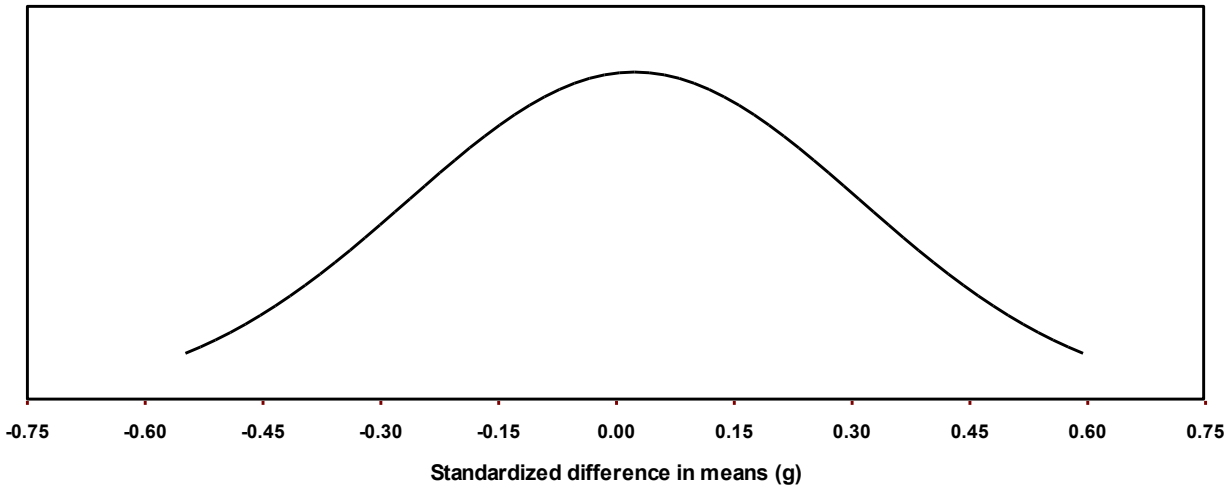
The first analysis replicates that of Wampold et al. (1997). In that study, studies were not classified into different types of therapy, and so no control group could be specified; accordingly,

directions could not meaningfully be assigned to the effect of any given comparison. Instead, effect directions were randomly assigned, which was achieved in this replication using a random number generator (Haahr, 2024). This necessarily creates an overall effect size that approximates zero. Consequently, the key measure in this analysis becomes the degree of homogeneity rather than the effect size: significant heterogeneity contradicts the hypotheses that all studies in the analysis share a common effect size (i.e., approximately zero), whereas no significant heterogeneity would not be inconsistent with the hypothesis of a shared effect size of approximately zero.

As expected, the mean effect size in this analysis showed no significant difference from zero using a criterion $\alpha = .05$, Hedges' $g = 0.02$, 95% CI [-0.03, 0.07], $Z = 0.88$, $p = .38$ (see Appendix I for a forest plot). However, adopting the standard criterion $\alpha = .10$, there was significant heterogeneity in this mean effect size, $Q = 2516.54$, $df = 158$, $p < .001$. The I^2 value indicates that 93.74% of the variance in observed effects reflects variance in true effects rather than sampling error. The standard deviation of true effect sizes was $\tau = 0.29$ in units of g, suggesting that ~68% of effect size differences have a magnitude of $\pm g = 0.29$. If we assume that the true effects are normally distributed (in units of Hedges' g), we can estimate that the true effect size in 95% of all comparable populations (i.e., the 95% prediction interval) varies between $g = -0.55$ and $g = 0.59$. (Borenstein, 2019, 2020; Borenstein et al., 2021; Borenstein et al., 2017; DerSimonian & Laird, 1986, 2015; Higgins, 2008; Higgins & Thompson, 2002; Higgins et al., 2003; Higgins & Thomas, 2019; IntHout et al., 2016). Figure 1 depicts the spread of randomly-signed effects about zero.

Figure 1

Distribution of Randomly-Signed Effect Sizes of Psychotherapies Compared to One Another



Note. The mean effect size is 0.02 (idealized value is 0.00). The effect size of 95% of comparisons is estimated to fall between -0.55 and 0.59.

Outliers. Six studies (Clark et al., 1994; Clark et al., 2006; Fairburn et al., 2015; Farahimanesh et al., 2021; Giesen-Bloo et al., 2006; Ost, 1988) had effects with 95% CIs that fell outside the 95% prediction interval for the average effect. The heterogeneity of the effects and the prediction interval were not deeply impacted by removing these studies, $Q(151) = 1757.91$, $p < .001$, $I^2 = 91.41\%$, $\tau = .24$, 95% PI [-0.44, 0.52]. Nevertheless, these studies were removed from the forest plot (Appendix I).

Summary. The results of the homogeneity test and metrics of dispersion are consistent with significant spread about zero, and do not replicate the findings of Wampold et al. (1997).

Overall Differences (Absolute Effect Sizes)

The second set of analyses addresses the question of differences between therapies. To approximate the overall magnitude of the difference of effect between various therapies, effect directions can be assigned all in the positive direction rather than randomly. Such an analysis of

absolute effect sizes cannot account for any actual negative effect sizes, and consequently likely overestimates the summary effect size. In previous analyses (e.g., Marcus et al., 2014), it has been interpreted as an upper limit on the average effect size of difference.

Across Timepoints and Comparisons

Combining across all timepoints and outcomes, the mean upper limit effect size is g = 0.27, 95% CI [0.23, 0.30]; this corresponds to a 57.6% chance that a participant selected randomly from the superior treatment will have favourable outcomes relative to a participant in the inferior treatment, where a 50% chance indicates no difference between treatments (Magnusson, 2023). In populations comparable to those in the analysis, these data are not consistent with a true mean upper limit effect size of zero (α = .05), Z = 14.40, p < .001. However, there is significant heterogeneity in this value (α = .10), inconsistent with the hypothesis that the true upper limit effect size is the same in all the included studies, Q = 1346.25, df = 158, p < .001. I^2 = 88.26%, indicating that 88% of variance in the observed effects reflects variance in true upper limit effects rather than sampling error. The estimated standard deviation of true effects was τ = 0.20, suggesting that the upper limit effect size falls between g = 0.06 and g = 0.46 in ~68% of comparable populations. The 95% prediction interval, g = -0.14–0.67, gives the range in which the upper limit true effect size falls in 95% of comparable populations.

Outliers. One study (Farahimanesh et al., 2021), had an effect size of g = 6.80, 95% CI [5.78, 7.82]; another (Giesen-Bloo et al., 2006) had an effect of g = 1.57, 95% CI [1.23, 1.89]. The 95% CIs for these effects fell outside the 95% prediction interval for the average effect. The results were not substantially impacted by removing these studies, g = 0.25, 95% CI [0.22, 0.29], Z = 14.74, p < .001, Q(156) = 1113.73, p < .001, I^2 = 85.99%, τ = .18, 95% PI [-0.11, 0.62].

Nevertheless, these studies were removed for all subsequent analyses. The forest plot of the effect sizes is found in Figure K1 (Appendix J).

Meta-Regression. A meta-regression was undertaken to determine the extent to which the variance in studies could be explained by differences between studies that compared treatments within a therapy family (e.g., CBT vs. CBT) as opposed to those comparing between families (e.g., CBT vs. IPT). Other moderators were the disorder of the population, the difference in researcher allegiance between comparators (absolute values, to correspond to the absolute value effect sizes), and study risk of bias per the Revised Cochrane Risk-of-Bias tool (RoB2; Sterne et al., 2019). Disorder categories were combined if they had less than 10 effects per covariate. The restricted maximum likelihood method was used to calculate the variance, τ^2, and the Knapp-Hartung adjustment was used for computing confidence intervals and *p*-values (Knapp & Hartung, 2003).

The test of the model containing all these covariates, $F(8, 154) = 1.82$, $p = .08$, is consistent with the hypothesis that none of these covariates explain any variation in absolute effect size. The R^2 analogue = .06, suggesting that the model can explain only 6% of the true variance. The goodness-of-fit test suggests that there is significant residual variance about the regression line, $\tau^2 = 0.04$, $Q(154) = 984.87$, $p < .001$, and that $I^2 = 84.36\%$ of the residual variance represents variance in true effects rather than sampling error. These results suggest that there is substantial true variance in the effects that is unexplained by these moderators.

Despite the non-significant test of the model, researcher allegiance was an individually significant predictor of the absolute effect size ($\alpha = .05$) when holding the other covariates constant. Every unit of difference in researcher allegiance score between one comparator and another score was associated with a $g = 0.04$ increase in absolute effect size, 95% CI [0.00,

0.07], $t(154) = 2.23$, $p = .03$. Disorder was not a significant independent predictor of effect size overall, $F(5, 154) = 2.18$, $p = .06$. However, individually, the absolute effect size was higher in the anxiety disorder population than in the depressive disorder population by $g = 0.14$, 95% CI [0.04, 0.24], $t(154) = 2.75$, $p = .01$, and in the trauma disorder population by $g = 0.14$, 95% CI [0.03, 0.26], $t(154) = 2.49$, $p = .01$. This suggests that the relation between each of these specific disorders and effect size cannot be explained as a confound of the other covariates in the model. No significant differences were seen when the depressive disorder population was compared to the addictive disorder population, $g = 0.13$, 95% CI [-0.02, 0.28], the obsessive-compulsive disorder population, $g = 0.02$, 95% CI [-0.14, 0.18], or other disorders, $g = 0.11$, 95% CI [-0.01, 0.23].

When controlling for the other moderators, the effect size was not independently predicted by either within- vs. between-family comparisons, $g = 0.05$, 95% CI [-0.03, 0.12], $t(154) = 1.20$, $p = .23$, or by RoB2 score, $g = -0.02$, 95% CI [-0.12, 0.09], $t(154) = -0.31$, $p = .76$.

Adjusting the overall model such that it only contained disorder and allegiance as moderators resulted in a significant test of the model, $F(6, 156) = 2.16$, $p = .05$; however, the proportion of total between-study variance explained by the model did not change, $R^2 = .06$.

Including only the primary effects that were based on comparisons between therapy families resulted in a non-significant model that explained an estimated 0% of the variance, $F(6, 51) = 0.69$, $p = .66$, $R^2 = .00$. Disorder, allegiance, and RoB2 score were not significant predictors of effect size among between-family comparisons, nor was any disorder category individually predictive. Removing or adding covariates did not change this pattern.

Subgroup Analysis. A mixed-effect analysis was conducted to compare the upper limit effect sizes of between within- vs. between-therapy family comparisons. A common among-

study variance component was not assumed across subgroups (i.e., within-group estimates of the variance were not pooled), as *ex hypothesi*, there may be substantially lower variance among studies that conducted within-family comparisons relative to between. In addition, the analysis has sufficient (50+) studies per subgroup.

In line with the meta-regression, the subgroup analysis revealed there is negligible difference in the upper limit effect between within- and between-family comparisons: within-family g = 0.24, 95% CI [0.20, 0.28], Z = 12.13, p < .001, τ = .17, 95% prediction interval: [-0.10, 0.58]; between-family g = 0.28, 95% CI [0.22, 0.34], Z = 8.77, p < .001, τ = .22, 95% prediction interval: [-0.16, 0.72]. These effects do not differ significantly with a criterion α = .10, $Q(1) = 1.01$, p = .31.

A mixed-effect analysis examining the impact of disorder population on effect size for all comparisons—pooling within-group estimates of the variance—is illustrated in Table 2 below. The effects do not differ significantly with a criterion α = .10, $Q(5) = 7.59$, p = .18.

Table 2

Effect Sizes and Heterogeneity Across Timepoints and Measures for Absolute Effect Sizes, by Disorder

Disorder	*k*	*g*	95% CI	CLES	*Z*	*p*	95% PI	T
Addictive disorders	11	0.23	0.11–0.36	56.5%	3.72	< .001	-0.14–0.61	0.18
Anxiety disorders	52	0.29	0.23–0.34	58.1%	9.85	< .001	-0.08–0.65	0.18
Depressive disorders	32	0.19	0.11–0.26	55.3%	5.00	< .001	-0.18–0.55	0.18
Obsessive-compulsive/ related disorders	12	0.17	0.05–0.30	54.8%	2.79	0.01	-0.20–0.55	0.18
Trauma disorders	28	0.30	0.22–0.38	58.4%	7.47	< .001	-0.06–0.67	0.18
Other disorders	22	0.26	0.17–0.35	57.3%	5.51	< .001	-0.11–0.62	0.18
Total	157	0.25	0.22–0.29	57.0%	14.91	< .001	-0.11–0.62	0.18

Note. *k* = number of studies; *g* = Hedges' *g*; CI = confidence interval; CLES = Common Language Effect Size or probability of superiority; *τ* = estimate of standard deviation of true effects in *g* units; PI = prediction interval in *g* units. "Other disorders" includes studies from the eating disorder, neurodevelopmental disorder, personality disorder, sleep disorder, and somatic disorder populations.

Publication Bias Analysis. The effects included in this meta-analysis contain only absolute effect sizes (i.e., no effect size above zero was included), which renders the usual tests of publication bias inapplicable (e.g., Egger's test, Duval & Tweedie's trim-and-fill procedure). Instead, a cumulative meta-analysis was conducted, calculating the average effect size starting with the largest, most precise study and successively adding smaller and less precise studies. By comparing the average effect when including only the 50% of the studies that are the largest with

the average effect when also including the smaller, less precise studies, one can appraise the impact of these smaller and potentially biased studies on the overall effect.

The upper limit effect size when including only the largest 81 studies was g = 0.22, 95% CI [0.17, 0.26], Z = 9.75, p < .001. When compared to the effect size when also including the 81 smallest studies, g = 0.26, 95% CI [0.22, 0.29], Z = 14.83, p < .001, the impact is not substantial. This suggests that, if smaller studies with larger absolute effect sizes were overrepresented in the analysis, the impact of their inclusion does not noticeably change the results.

Summary. Collapsing across timepoints and outcome measures, the main findings for the meta-analyses of absolute value effect size include the following:

1. The overall effect was g = 0.25, 95% CI [0.22, 0.29],
2. There was substantial heterogeneity in this effect, I^2 = 85.99%, τ = .18, 95% PI [-0.11, 0.62].
3. In the meta-regression model, the moderators together explained only 6% of the true variance. The test of the model containing all of them, $F(8, 154) = 1.82$, p = .08, is consistent with no moderators explaining any variation in absolute effect size.
4. In the meta-regression, whether a comparison was between- or within-family did not predict effect size, g = 0.05, 95% CI [-0.03, 0.12], p = .23

a. In the subgroup analysis, not controlling for the other moderators, the effect within-families g = 0.24, 95% CI [0.20, 0.28] and between-families g = 0.28, 95% CI [0.22, 0.34], were not significantly different from one another, p = .31.

By Timepoint and Measure Type

The main analyses for this review distinguish among effects at both at different timepoints (at therapy termination vs. 6+ months follow-up) and different classes of measures

(i.e., symptom-specific measures related to the primary disorder in the sample; symptom-specific measures related to secondary disorder(s) in the sample; and global, non-symptom-specific measures). The effect size and heterogeneity measures are presented for each measure type at termination in Table 3, and for each measure type at follow-up in Table 4. Forests plots for each analysis can be found in Appendix J (Figures J2–J7).

Table 3

Effect Sizes and Heterogeneity Measures, Absolute Effect Sizes at Termination, Outliers Removed

Primary disorder measures, k = 150

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.27	0.23–0.31	57.6%	13.46	< .001

Q	*df*	*p*	I^2	*τ*	95% PI
453.91	149	< .001	67.17%	0.18	-0.09–0.63

Outliers Farahimanesh et al., 2021; Fonagy et al., 2020; Giesen-Bloo et al., 2006

Secondary disorder measures, k = 96

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.27	0.22–0.32	57.6%	10.75	< .001

Q	*df*	*p*	I^2	*τ*	95% PI
--	--	--	22.85%	0.11	0.05–0.49

Outliers Fairburn et al., 2015; Farahimanesh et al., 2021

Global measures, k = 42

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.30	0.23–0.37	58.4%	8.58	< .001

Q	*df*	*p*	I^2	*τ*	95% PI
--	--	--	19.01	0.12	0.05–0.55

Outliers Fairburn et al., 2015

Table 3

Effect Sizes and Heterogeneity Measures, Absolute Effect Sizes at Termination, Outliers Removed

Note. k = number of studies; CI = confidence interval; CLES = Common Language Effect Size or probability of superiority; I^2 = proportion of overall variance representing variance in true effects; τ = estimate of standard deviation of true effects in g units; PI = prediction interval in g units.

At termination, the 95% confidence range for the mean upper limit effect size varied from $g = 0.22$ at the lowest (secondary symptom measures) to $g = 0.37$ at the highest (global measures). A psychotherapy client could therefore expect that by being in a superior treatment, on average, their probability of better outcomes than the alternative at termination improves marginally over chance by between 6.2% and 10.3% (depending on the outcome in question). The effect for primary measures had significant heterogeneity, indicating that the effect size cannot be interpreted straightforwardly: for primary symptom measures, the 95% PIs cross zero, suggesting that there will be a significant difference between treatments in some relevant populations and no difference in others. For secondary symptom and global measures, the analysis was underpowered to determine whether there was significant heterogeneity and are therefore not reported here. The 95% PIs of these effects do not cross zero, suggesting that in 95% of all relevant populations, there will be differences between treatments of at least $g = 0.05$.

Table 4

Effect Sizes and Heterogeneity Measures, Absolute Effect Sizes at Follow-Up, Outliers Removed

Primary disorder measures, *k* = 89

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.24	0.19–0.28	56.7%	10.14	< .001

Q	*df*	*p*	I^2	*τ*	95% PI
288.44	88	< .001	69.49%	0.16	-0.09–0.56

Outliers n/a

Secondary disorder measures, *k* = 96

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.22	0.17–0.27	56.2%	8.63	< .001

Q	*df*	*p*	I^2	*τ*	95% PI
--	--	--	17.82%	.08	0.06–0.39

Outliers Fairburn et al., 2015; Ost, 1988

Global measures, *k* = 74

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.19	0.13–0.24	55.3%	6.58	< .001

Q	*df*	*p*	I^2	*τ*	95% PI
--	--	--	35.90%	0.10	-0.02–0.39

Outliers Fairburn et al., 2015

Note. *k* = number of studies; CI = confidence interval; CLES = Common Language Effect Size or probability of superiority; I^2 = proportion of overall variance representing variance in true effects; *τ* = estimate of standard deviation of true effects in *g* units; PI = prediction interval in *g* units.

At follow-up, the 95% confidence range for the mean upper limit effect size varied from $g = 0.19$ at the lowest (global measures) to $g = 0.24$ at the highest (primary symptom measures). A psychotherapy client could therefore expect that by being in a superior treatment, on average, their probability of better outcomes than the alternative at termination improves marginally over chance by between 5.3% and 6.7% (depending on the outcome in question). The effect for primary symptom measures had significant heterogeneity, indicating that the effect sizes cannot be interpreted straightforwardly: its 95% PI crossed zero, suggesting that there will be a significant difference between treatments in some relevant populations and no difference in others. The 95% PI for global measures also crossed zero, whereas the 95% PI for secondary symptom measures did not; however, for both of these analyses, there was insufficient power for the *Q* test to detect heterogeneity (and so these results are not reported here).

Because the PI for secondary symptom measures did not include zero, the average upper limit effect size of $g = 0.22$, 95% CI [0.17, 0.27], could be interpreted straightforwardly, with no reason to believe that it would vary substantially as a function of moderating variables. A client could therefore expect that by being in a superior treatment, on average, their probability of better secondary disorder outcomes at follow-up improves marginally over chance by between 4.8% and 7.6%. Furthermore, in 95% of all relevant populations, there will be differences between treatments of at least $g = 0.06$ for secondary symptom measures at follow-up.

Meta-Regression and Subgroup Analyses

The analysis of primary symptom measures at termination revealed a large degree of heterogeneity and a moderate-to-large value of I^2; the *Q* test was sufficiently powered at $\beta = .80$ and $\alpha = .05$ to detect this degree of heterogeneity, and indeed, the *Q* test was inconsistent with the null hypothesis that the true upper limit effect size is the same in all the included studies.

However, for secondary disorder and global measures, the degree of heterogeneity and the value of I^2 were both small, and the Q test was underpowered to reject the null hypothesis. The same pattern was true at follow-up. Consequently, moderator and subgroup analyses were undertaken for primary symptom measures only.

As before, meta-regression was used to test the impact of four planned moderators: within- vs. between-family comparisons, disorder population, difference in researcher allegiance between comparators, and Risk of Bias 2 score. As before, disorder categories were combined if there were less than 10 effects per covariate. Sensitivity analysis revealed that outlying studies (i.e., those outside the 95% prediction interval of the overall effect) had a disproportionate impact on the results of the meta-regression. All analyses below present the results with outliers removed.

Primary Disorder-Specific Outcomes, Termination. The test of the model containing all the planned moderators, $F(8, 147) = 2.31$, $p = .02$, suggests that at least one of the moderators explains some variation in effect size ($\alpha = .05$). The R^2 analogue = .14, which implies that the model explains 14% of the overall variance. The goodness-of-fit test of the meta-regression is consistent with significant residual variance about the regression line, $\tau^2 = 0.03$, $Q(147) = 417.45$, $p < .001$, and that $I^2 = 64.79\%$ of the residual variance represents variance in true effects rather than sampling error. These results suggest that there is substantial true variance in the effects that is unexplained by any of these moderators.

Disorder was a significant predictor of the effect size when holding other moderators constant, $F(5, 147) = 2.71$, $p = .02$. Specifically, relative to the average of comparisons in the depressive disorder population, the average upper limit effect size of comparisons were higher in populations with: anxiety disorders by $g = 0.18$, 95% CI [0.06, 0.30], $t(147) = 2.71$, $p = .004$;

trauma disorders by $g = 0.22$, 95% CI [0.09, 0.36], $t(147) = 3.26$, $p = .001$; and other disorders by $g = 0.17$, 95% CI [0.02, 0.32], $t(147) = 2.24$, $p = .03$. The difference in average effects when comparing studies in the depressive disorder and the addictive disorder population was not significant, $g = 0.16$, 95% CI [-0.02, 0.34], $t(147) = 1.74$, $p = .08$, nor was it in the obsessive-compulsive disorder population, $g = 0.06$, 95% CI [-0.13, 0.25], $t(147) = 0.64$, $p = .52$.

Researcher allegiance was also a significant independent predictor of effect size when holding the other moderating variables constant: for every unit increase in the difference in allegiance scores between two comparators, the effect size increased by $g = 0.05$, 95% CI [0.01, 0.09], $t(147) = 2.79$, $p = .01$. When controlling for the other moderators, the effect size was not independently predicted by either within- vs. between-family comparisons, $g = 0.06$, 95% CI [-0.03, 0.14], $t(147) = 1.23$, $p = .22$, or by RoB2 score, $g = 0.01$, 95% CI [-0.12, 0.14], $t(147) = .15$, $p = .88$.

Table 5 illustrates the impact of disorder population on effect size per a subgroup analysis with a mixed effect model—pooling within-group estimates of the variance and combining subgroup using a fixed-effects model. Effect sizes were significantly ($\alpha = .10$) different from one another, $Q(5) = 9.42$, $p = .09$.

Table 5

Effect sizes and Heterogeneity Across Measures at Termination for Absolute Effect Sizes, By Disorder

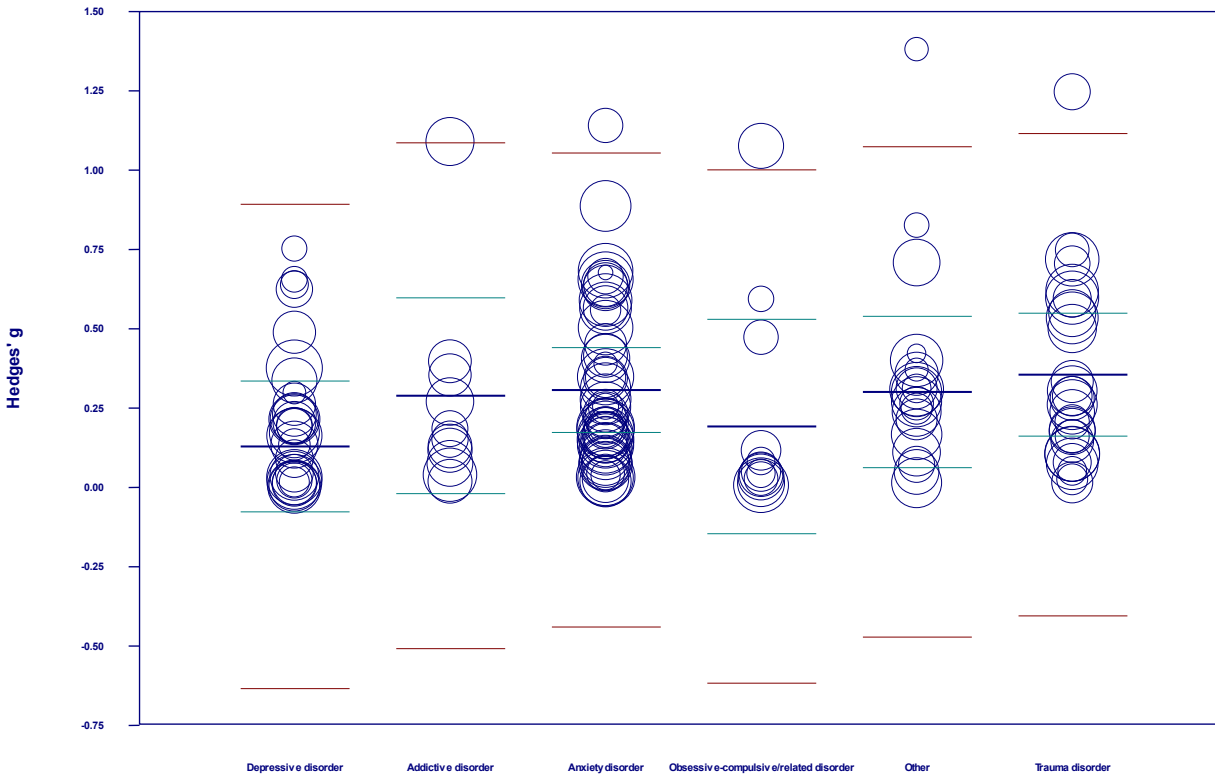
Disorder	*g*	95% CI	CLES	*Z*	*p*	95% PI	τ
Addictive disorders	0.22	0.07–0.37	56.2%	2.83	< .001	-0.17–0.60	0.18
Anxiety disorders	0.29	0.23–0.36	58.1%	9.09	< .001	-0.07–0.65	0.18
Depressive disorders	0.18	0.09–0.27	55.1%	3.94	< .01	-0.19–0.54	0.18
Obsessive-compulsive/ related disorders	0.20	0.04–0.35	54.8%	2.54	.01	-0.19–0.58	0.18
Trauma disorders	0.35	0.26–0.45	59.8%	7.55	< .001	-0.01–0.72	0.18
Other disorders	0.29	0.18–0.40	58.1%	5.16	< .001	-0.08–0.66	0.18
Total	0.27	0.23–0.31	57.0%	14.44	< .001	-0.09–0.63	0.18

Note. g = Hedges' *g*; CI = confidence interval; CLES = Common Language Effect Size or probability of superiority; *τ* = estimate of standard deviation of true effects in *g* units; PI = prediction interval in *g* units. "Other disorders" includes studies from the eating disorder, neurodevelopmental disorder, personality disorder, sleep disorder, and somatic disorder populations.

Figure 2 depicts the scatter plot of studies' effect sizes at termination as a function of disorder, and Figure 3 plots effect sizes against researcher allegiance score.

Figure 2

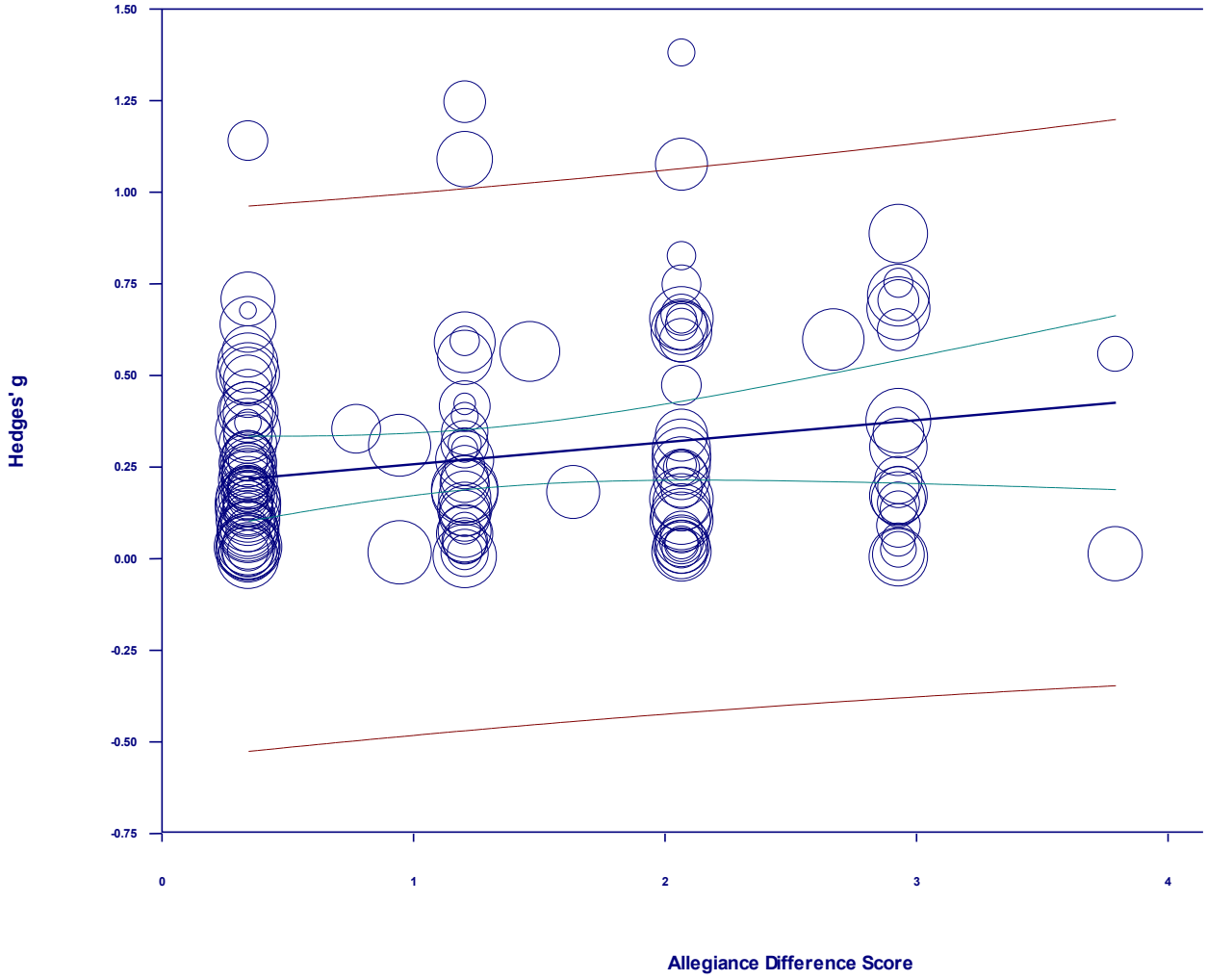
Association Between Absolute Effect Size and Disorder Category, Primary Disorder Measures at Termination



Note. Each dot represents a primary RCT and is proportional to the sample size of the study. Blue lines represent the mean effect size. Green lines represent the bounds of the 95% confidence interval. Red lines represent the bounds of the 95% prediction interval.

Figure 3

Association Between Absolute Effect Size and Researcher Allegiance Difference Score, Primary Disorder Measures at Termination



Note. Regression of absolute effect size against the absolute difference between conditions in their researcher allegiance scores (higher scores represent greater discrepancy in allegiance bias). Each dot represents a primary RCT and is proportional to the sample size of the study. The regression line is depicted in blue. Green lines represent the bounds of the 95% confidence interval. Red lines represent the bounds of the 95% prediction interval.

Primary Disorder-Specific Outcomes, Follow-Up. The test of the model containing the planned moderators was marginally significant, $F(6, 86) = 2.20$, $p = .0509$, arguably consistent

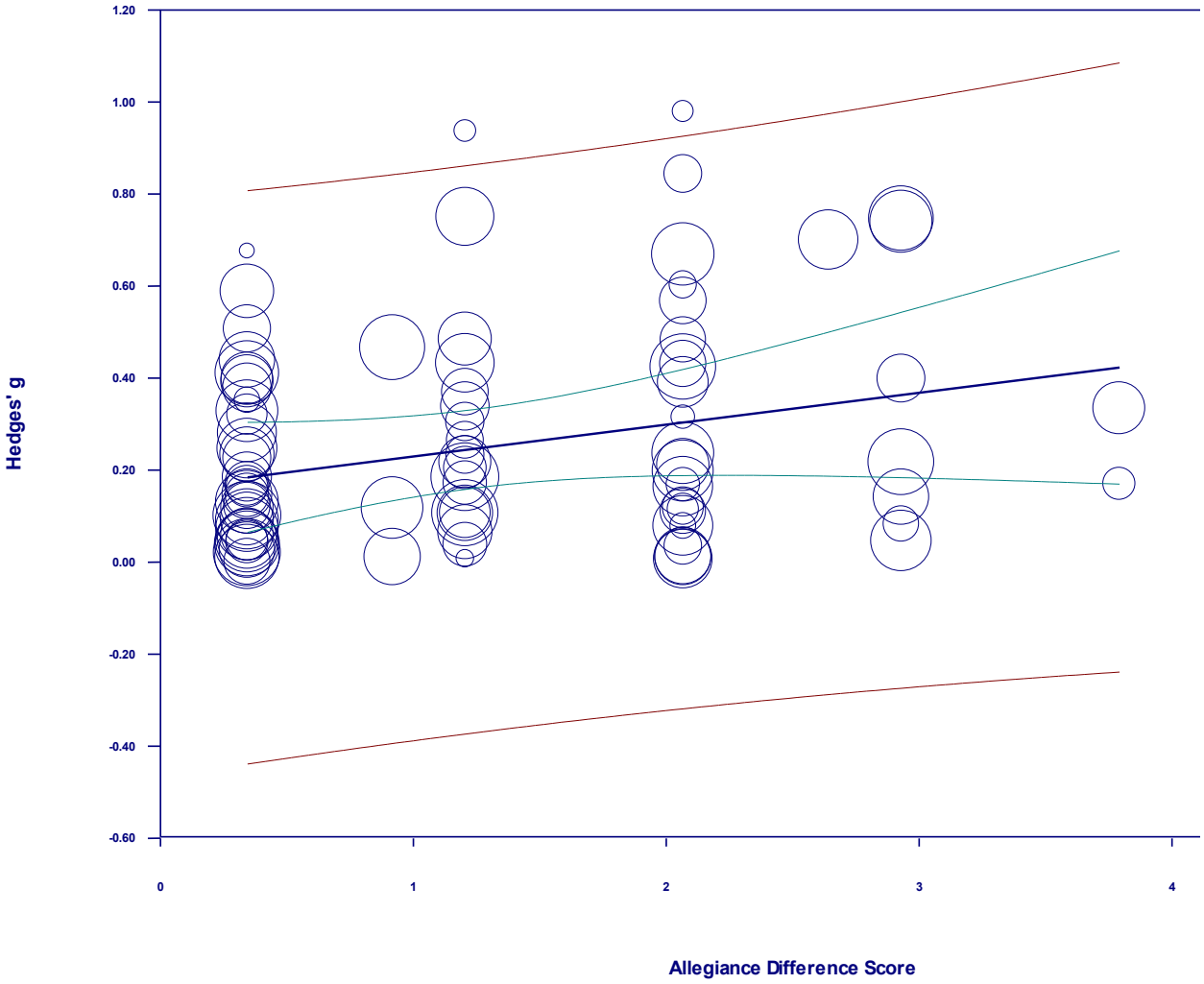
with at least one of the moderators explaining variation in the summary effect size ($\alpha = .05$). The R^2 analogue = .16, which implies that the model explains 16% of the overall variance. The goodness-of-fit test is consistent with significant residual variance about the regression line, $\tau^2 = 0.03$, $Q(86) = 275.12$, $p < .001$. $I^2 = 68.74\%$ of the residual variance represents variance in true effects rather than sampling error.

The effect size was significantly predicted by researcher allegiance score when holding the other moderators constant: for unit increase in allegiance score, effect size changed by $g = -0.06$, 95% CI [0.02, 0.10], $t(86) = 2.79$, $p = .01$. Disorder did not predict effect size overall, $F(3, 86) = 1.1$, $p = .17$; however, individually, the absolute effect size was higher in the trauma disorder population by $g = 0.19$, 95% CI [0.02, 0.36], $t(86) = 2.21$, $p = .03$. Effect size was not significantly independently predicted by within- vs. between-family comparisons, $g = -0.001$, 95% CI [-0.10 , 0.10], $t(86) = -0.02$, $p = .98$; or by RoB2 score, $g = -0.01$, 95% CI [-0.17, 0.16], $t(86) = -0.08$, $p = .94$.

Figure 4 depicts the scatter plots of effect sizes for primary measures at follow-up against researcher allegiance score.

Figure 4

Association Between Absolute Effect Size and Researcher Allegiance Difference Score, Primary Disorder Measures at Follow-Up



Note. Regression of absolute effect size against the absolute difference between conditions in their researcher allegiance scores (higher scores represent greater discrepancy in allegiance bias). Each dot represents a primary RCT and is proportional to the sample size of the study. The regression line is depicted in blue. Green lines represent the bounds of the 95% confidence interval. Red lines represent the bounds of the 95% prediction interval.

Summary. When accounting for timepoint and outcome type, and after addressing outliers, the main findings for the meta-analyses of absolute value effect size include the following:

1. At termination,
    a. The overall effects were
        i. g = 0.27, 95% CI [0.23, 0.31] for primary disorder measures, $\boldsymbol{\tau}$ = .18, 95% prediction interval [-0.09, 0.63]
        ii. g = 0.27, 95% CI [0.22, 0.32] for secondary disorder measures, $\boldsymbol{\tau}$ = .11, 95% prediction interval [0.05, 0.49]
        iii. g = 0.30, 95% CI [0.23, 0.37] for global measures, $\boldsymbol{\tau}$ = .12, 95% prediction interval [0.05, 0.55]
    b. For primary disorders, the meta-regression model containing all moderators was significant, p = .02, and explained 14% of the variance.
        i. Holding the other moderators constant, disorder was a significant predictor of effect size, p = .02.
            1. Relative to effect sizes in RCTs in the depressive disorder population,
                a. those in the anxiety disorder population had a higher effect size by g = 0.18, 95% CI [0.06, 0.30].
                b. in the trauma population by g = 0.22, 95% CI [0.09, 0.36]
                c. in “other” disorders by g = 0.17, 95% CI [0.02, 0.32]
            2. In the subgroup analysis, not controlling for other moderators,
                a. depression g = 0.18, 95% CI [0.09, 0.27]

b. anxiety g = 0.29, 95% CI [0.23, 0.36]

c. trauma g = 0.35, 95% CI [0.26, 0.45]

d. "other" g = 0.29, 95% CI [0.18, 0.40]

ii. Holding the other moderators constant, researcher allegiance was a significant predictor of effect size, p = .01. For every unit increase in the difference in allegiance scores between two comparators, the effect size increased by g = 0.05, 95% CI [0.01, 0.09].

2. At follow-up,

a. The overall effects were

i. g = 0.24, 95% CI [0.19, 0.28] for primary disorder measures, $\boldsymbol{\tau}$ = .16, 95% prediction interval [-0.09, 0.56]

ii. g = 0.22, 95% CI [0.17, 0.27] for secondary disorder measures, $\boldsymbol{\tau}$ = .08, 95% prediction interval [0.06, 0.39]

iii. g = 0.19, 95% CI [0.23, 0.37] for global measures, $\boldsymbol{\tau}$ = .10, 95% prediction interval [-0.02, 0.39]

b. For primary disorders, the meta-regression model containing all moderators was marginally significant, $F(6, 86) = 2.20$, p = .0509, and explained 16% of the variance.

i. Holding the other moderators constant, researcher allegiance was a significant predictor of effect size, p = .01. For every unit increase in the difference in allegiance scores between two comparators, the effect size decreased, g = -0.06, 95% CI [-0.02, -0.10].

Cognitive Contrast (Standard Effect Sizes Comparisons)

To compare therapy families using standard meta-analytic techniques and producing standard meta-analytic effect sizes, a control group must be specified. For this third set of analyses, CBT was chosen as the control condition, as CBT was the therapy family with the greatest number of studies in which it was a comparator.

Across Timepoints and Comparisons

Averaging across timepoints and measures, the mean effect size of CBT vs. other therapy families was $g = 0.11$ with a 95% confidence interval [0.02, 0.21], $Z = 2.38$, $p = .02$. There is significant heterogeneity in the studies, $Q(52) = 1005.43$, $p < .001$. The standard deviation of true effect sizes, τ, is $g = 0.33$, and the 95% prediction interval (the range in which the true effect size falls in 95% of all comparable populations, assuming the effects are normally distributed) is $g = -0.56$ to $g = 0.79$.

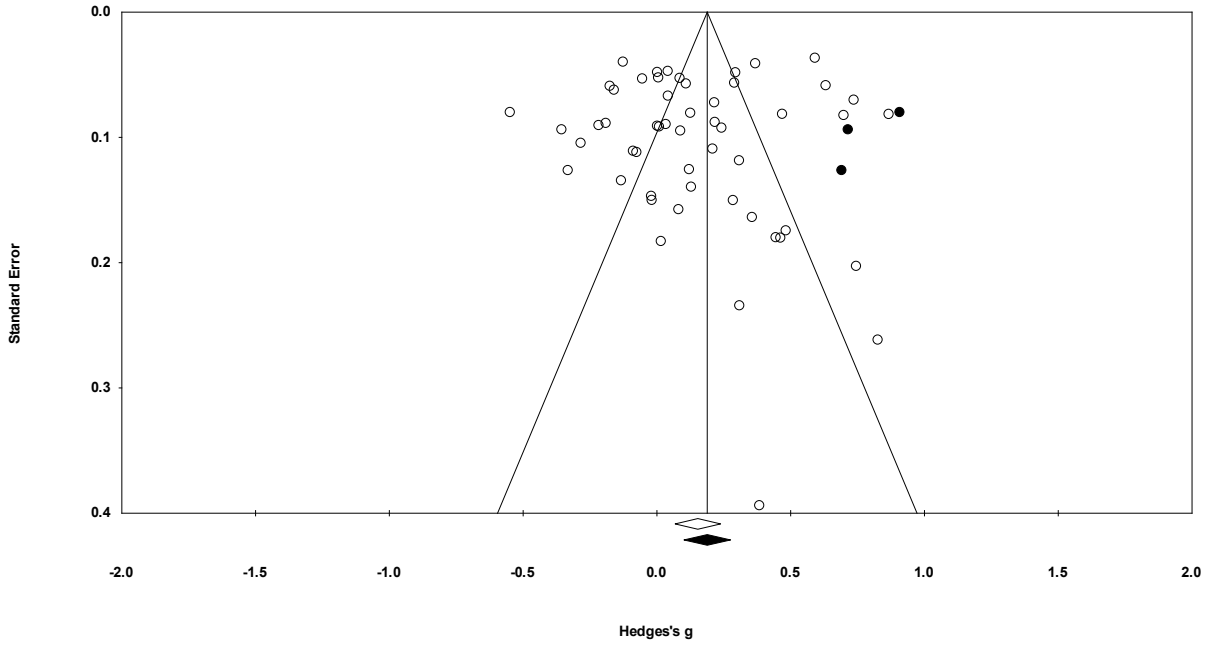
Outliers. Two CBT-CBT comparisons (Clark et al., 2006; Farahimanesh et al., 2021) and two CBT-other comparisons (Capezzani et al., 2013; Fonagy et al., 2020) had a 95% confidence interval that did not overlap with the 95% prediction interval of their respective summary effect sizes. Removing these studies, the mean effect size of CBT vs. other therapy families was $g = 0.15$ with a 95% confidence interval [0.06, 0.24], $Z = 3.30$, $p = .001$. There was significant heterogeneity in the studies, $Q(50) = 806.55$, $p < .001$, $I^2 = 93.80\%$. The standard deviation of true effect sizes, τ, is $g = 0.30$, and the 95% prediction interval (the range in which the true effect size falls in 95% of all comparable populations, assuming the effects are normally distributed) is $g = -0.45$ to $g = 0.76$. The outlying studies were removed from all further analyses. The forest plot of the effect sizes is found in Figure K1 (Appendix K).

Publication Bias Analyses. Per Egger et al.'s (1997) test of the intercept, the effect of bias, B = -0.89, 95% CI [-3.56, 1.78], $t(49) = .67$, $p = .25$, is not consistent with a small-study effect or publication bias.

Using Duval and Tweedie's (2000) trim-and-fill procedure—trimming studies per the fixed effect model and filling per the random effects model—the effect size was $g = 0.15$, 95% CI [0.06, 0.24]. This is unchanged from the observed value, inconsistent with any evidence of a small-study effect or publication bias. Indeed, there was some evidence of a "large-study effect," whereby there were relatively fewer large studies with large effects in favour of CBT, and relatively more large studies with small effects, than expected. Setting the trim-and-fill procedure to correct for this asymmetry produced an increased effect of $g = 0.19$, 95% CI [0.10, 0.28].

Figure 5

Funnel Plot of Standard Error by Hedges g *For CBT vs. Other Therapies, Collapsed Across Timepoints and Outcomes*



Note. White dots represent individual studies. Black dots represent the effect of "studies" imputed by the trim-and-fill procedure.

Given that data collection concluded by 2021, analyses were undertaken to estimate the potential impact of missing recent studies. Specifically, meta-regressions of year on effect size were undertaken for the Cognitive Contrast analysis, examining the impact for studies 1980-2021 as well as 2015-2021. Additionally, an Orwin's Failsafe N analysis was undertaken to determine how many studies would be required to change the conclusions of the Cognitive Contrast analysis. The results of this analysis, presented in Appendix M, suggest that the inclusion of recent studies would be very unlikely to alter the findings of the Cognitive Contrast analysis.

Meta-Regression. A meta-regression was conducted to determine whether comparator predicted effect size when controlling for disorder, RoB2 score, and differences in allegiance. To ensure that each covariate contained at least 10 comparisons, disorders with fewer than this number (viz., addictive, eating, neurodevelopmental, personality, or trauma disorders; there were no comparisons in the obsessive-compulsive, sleep, or somatic disorders categories) were collapsed into one category. Similarly, CBT vs. EMDR comparisons were not included as a covariate in the main analysis as there were only six comparisons.

The test of the model containing the four planned moderators was not consistent with the hypothesis that at least one moderator explained variation in effect size, $F(7, 41) = .078$, $p = .61$. The R^2 analogue = .06, suggesting that the model can explain 6% of the true variance in effects. The goodness-of-fit tests suggests that there is significant residual variance about the regression line, $\tau^2 = 0.07$, $Q(41) = 505.25$, $p < .001$; $I^2 = 91.87\%$ of the residual variance represents true variance in the effects rather than sampling error. No covariate was an individually significant independent predictor of effect size. These results suggest that there is substantial true variance in the effects that is unexplained by these moderators. Removing RoB2 score and allegiance

score difference improved the value of $R^2 = .14$, but the test of the model remained inconsistent with the idea that any moderator significantly predicted effect size, $F(5, 43) = 1.04$, $p = .41$.

As sensitivity analyses, the meta-regression was re-run first with CBT vs. EMDR comparisons included but underpowered disorders still collapsed, and then again with all underpowered covariates included (i.e., addictive, eating, neurodevelopmental, personality, and trauma disorders; CBT vs. EMDR comparisons). In the former analysis, no substantial differences emerged: the test of the model was not significant, $F(9, 44) = 1.42$, $p = .21$, $R^2 = .12$, and no moderator significantly independently predicted effect size. The same was largely true of the latter analysis: test of the model yielded $F(12, 41) = 1.25$, $p = .29$, $R^2 = .14$. Comparator, disorder, RoB2, and allegiance did not predict effect size. However, anxiety disorder emerged as an individually significant predictor of effect size when all other moderators were held constant, $g = 0.31$, 95% CI [0.02, 0.50], $t(41) = 2.16$, $p = .04$. This result is inconsistent with the idea that the relationship between anxiety disorders and effect size is confounded by other covariates in the model.

Lastly, CBT-CBT comparisons, using absolute effect sizes, were included in the meta-regression to determine the extent to which effect sizes differed when comparing one type of CBT to another vs. comparing a CBT to another therapy family. As before, CBT vs. EMDR comparisons were not included as a covariate in the main analysis as there were insufficient comparisons, and disorder categories with fewer than 10 comparisons (viz., eating, neurodevelopmental, personality, sleep, and somatic disorders) were collapsed. When disorder, risk of bias, and allegiance were held constant, comparator did not significantly predict the effect size independently, $F(4, 135) = 0.54$, $p = .71$. No comparator was an individually significant predictor of effect size. The test of the model was not significant, $F(8, 138) = 1.87$, $p = .07$, and

the value of $R^2 = .09$, suggesting that these moderators explain only 9% of the true variance in effects. The goodness-of-fit test indicates that there was significant unexplained variance, $\tau^2 = 0.04$, $Q(135) = 1021.83$, $p < .001$, and $I^2 = 86.79\%$ of the residual variance represents variance in the true effects. In all, these results suggest substantial true variance is unexplained by the included moderators.

Subgroup Analysis. A mixed-effect analysis was conducted to compare how effect sizes varied as a function of the comparator to CBT when not controlling for other potential moderators. As before, the absolute value of the effect size was assumed for CBT-CBT comparisons, whereas for all other analyses, a positive effect size indicated a larger effect for CBT whereas a negative effect favoured the comparator. A common among-study variance component was assumed across subgroups (i.e., pooled within-group estimates of the variance), as only one subgroup (CBT-CBT comparisons) had sufficient (20+) studies to assume otherwise. Results are presented in Table 6 below.

Table 6

Effect Sizes and Heterogeneity Measures Across Timepoints and Measures Relative to CBT, By Comparator

	k	*g*	95% CI	CLES	*Z*	*p*	95% PI	T
CBT	98	0.24	0.19–0.29	56.7%	10.05	< .001	-0.17–0.65	0.20
EHT	11	0.16	0.02–0.29	54.5%	2.23	.03	-0.27–0.58	0.20
EMDR	5	-0.18	-0.38–0.03	44.9%	-1.67	.09	-0.63–0.29	0.20
Integr.	12	0.26	0.13–0.39	57.3%	3.82	< .001	-0.26–0.59	0.20
IPT	14	0.17	0.05–0.29	54.8%	2.70	.01	-0.25–0.59	0.20
PDT	12	0.12	-0.01–0.25	53.4%	1.78	.08	-0.31–0.54	0.20
Combined	51	0.15	0.06–0.23	54.2%	3.37	.001	-0.46–.076	0.30
Overall	143	0.20	0.16–0.24	57.9%	10.71	<.001	-0.20–0.60	0.20

k = number of comparisons; *g* = Hedges' *g*; CI = confidence interval; CLES = Common Language Effect Size or probability of superiority; *τ* = estimate of standard deviation of true effects in *g* units; PI = prediction interval in *g* units. Positive effect sizes favour CBT. Combined = all comparators except CBT. Overall = all comparators. Estimates of *g*, *Z*, CI and PI for CBT were assigned positive signs, but all signs can be reversed arbitrarily.

The effect sizes in Table 6 are significantly different from one another, $Q(5) = 18.34$, $p = .003$, implying (on the basis of the 95% CIs) that CBT was superior when compared to EHT, IPT, and integrative therapies, and that there was no significant difference when compared to EMDR or PDT. However, no effect size meaningfully exceeded the magnitude of effect size of one type of CBT compared to another (compare: 56.6% probability of superiority for one CBT vs. another; 57.3% probability of superiority of CBT vs. integrative). Moreover, the combined effect of all comparisons between CBT and another therapy family is lower than the overall effect when including CBT vs. CBT comparisons. These results are inconsistent with the notion that differences between therapy families exceed differences within therapy families.

Summary. For the meta-analyses of effect sizes when comparing various therapy families to CBT (i.e., the effect sizes when classifying by treatments), and collapsing across timepoints and measurements, the main findings after accounting for outliers include the following:

1. The overall effect size between CBT and other families was $g = 0.15$, 95% CI [0.06, 0.24], $p = .001$.
2. There was substantial heterogeneity in this effect, $p < .001$, $I^2 = 93.80$, $\tau = 0.30$, 95% prediction interval [-0.46, 0.76].

3. In the meta-regression model, the moderators together explained only 6% of the true variance. The test of the model containing all the moderators, $F(7, 41) = .078$, $p = .61$, is consistent with none of them explaining any variation in absolute effect size.
4. In the subgroup analysis (i.e., not controlling for the other moderators), CBT had significantly better outcomes than EHT, $g = 0.16$, 95% CI [0.02, 0.29]; Integrative therapies, $g = 0.26$, 95% CI [0.13, 0.39]; and IPT, $g = 0.17$, 95% CI [0.05, 0.29]. However, the absolute value effect size of one CBT intervention relative to another, $g = 0.24$, 95% CI [0.19, 0.29], was also significant and was not meaningfully different from CBT versus other families (at best, <1% marginal increase in the probability of superiority).

Termination

The average effect of CBT vs. other therapy families at termination, along with measures of heterogeneity, are presented in Table 7. As before, outliers (with 95% CIs that fall outside the 95% PI for the summary effect) were excluded. See Figures K2–K4 in Appendix K for forest plots.

Table 7

Effect Sizes and Heterogeneity Measures, Between-Family Effect Sizes at Termination, Outliers Removed

Primary disorder measures, $k = 50$

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.19	0.09–0.29	55.3%	3.67	< .001

Q	*df*	*p*	I^2	τ	95% PI
323.04	49	< .001	84.83	0.31	-0.45–0.83

Outliers Capezzani et al., 2013; Fonagy et al., 2020

Table 7

Effect Sizes and Heterogeneity Measures, Between-Family Effect Sizes at Termination, Outliers Removed

Secondary disorder measures, k = 26

Hedges' g	95% CI	CLES	Z	p
0.09	-0.08–0.25	52.5%	--	--

Q	df	p	I^2	τ	95% PI
120.11	25	< .001	79.19	0.35	-0.66–0.83

Outliers Budney et al., 2000; Fairburn et al., 2015

Global measures, k = 29

Hedges' g	95% CI	CLES	Z	p
0.05	-0.06–0.16	51.4%	--	--

Q	df	p	I^2	τ	95% PI
121.50	28	< .001	76.96	0.24	-0.45–0.55

Outliers Fairburn et al., 2015

Note. Effect sizes of CBT vs. other therapy families; positive effect sizes favour CBT. k = number of studies; CI = confidence interval; CLES = Common Language Effect Size or probability of superiority; I^2 = proportion of overall variance representing variance in true effects; τ = estimate of standard deviation of true effects in g units; PI = prediction interval in g units.

At termination, the 95% confidence range for the average effect of CBT vs. alternative therapy varied from g = -0.08 at the lowest (secondary symptom measures) to g = 0.29 at the highest (primary symptom measures). A psychotherapy client could expect that by enrolling in CBT, on average, their chance of a better outcome than by enrolling in an alternative could be as low as 2.3% worse than chance and as high as an 8.1% improvement over chance (depending on the outcome in question). The effect for primary disorder was significant (α = .05); the analysis

was insufficiently powered to determine if the average effect was significantly different from zero for secondary disorder and global measures. However, the 95% CI included zero for both secondary and global measures. All effects had significant heterogeneity, indicating that the effect sizes cannot be interpreted straightforwardly: in 95% of relevant populations, the effect will favour CBT in some circumstances (by as much as $g = 0.83$) and the alternative therapy in others (by as much as $g = 0.66$).

Meta-Regression and Subgroup Analyses. At termination, the analyses of all three outcomes (primary disorder, secondary disorder, and global) revealed large quantities of heterogeneity and large values of I^2; the Q tests also detected significant heterogeneity and were sufficiently powered to do so at $\beta = .80$ and $\alpha = .10$. However, for secondary and global measures, the meta-regression was itself underpowered: for the four planned moderators, a minimum of 40 effects would be required, and both analyses had fewer. Consequently, the results of the meta-regression were reported only for primary disorder outcomes.

Primary Disorder-Specific Outcomes. To assess which factors may explain the variation in effect size, a meta-regression was undertaken including the following potential moderators: therapy family (of the comparator to CBT); disorder; the study's risk of bias (RoB2 tool score); and difference in researcher allegiance score between the two comparators. Outliers were excluded and disorder categories with fewer than 10 comparisons per category (viz., addictive, eating, neurodevelopmental, personality, and trauma disorders; no comparisons were available for obsessive-compulsive, sleep, or somatic disorders) were collapsed. Initially, EMDR vs. CBT comparisons were excluded as a covariate due to having fewer than 10 effects. The meta-regression was then repeated with EMDR comparisons included as a sensitivity analysis (and

including trauma disorders as a separate category, as there were 10 effects when including EMDR comparisons).

The test of the initial meta-regression model was not significant, ($\alpha = .05$), $F(7, 40) = 1.41$, $p = .22$, suggesting that none of the variation in effect size can confidently be ascribed to any of the moderators in the model. The R^2 analogue = .17, which implies that the model explains 17% of the overall variance. Replacing RoB2 score with an ad-hoc moderator—year of publication—increased the proportion of variance explained by the model, $R^2 = .23$, but the test of the model remained non-significant, $F(7, 40) = 1.79$, $p = .12$. The goodness-of-fit test for this adjusted meta-regression is consistent with significant residual variance about the regression line, $\tau^2 = 0.06$, $Q(40) = 166.26$, $p < .001$, and that $I^2 = 75.94\%$ of the residual variance represents variance in true effects rather than sampling error. These results suggest that there is substantial true variance in the effects that is unexplained by any of these moderators. No covariate individually significantly explained variation in effect size when accounting for the other variables in the model.

Including EMDR comparisons in the model containing comparator, disorder, allegiance, and year of publication yielded a substantially different pattern of results. The test of this model was significant, $F(9, 43) = 1.94$, $p = .02$, and $R^2 = .31$ (i.e., the model accounted for 31% of the true variance in effects). The results of the heterogeneity tests remained essentially unchanged, $\tau^2 = 0.07$, $Q(43) = 183.27$, $p < .001$, $I^2 = 76.54\%$. No moderator emerged as a significant independent predictor of effect size: comparator, $F(4, 43) = 2.28$, $p = .08$; disorder, $F(3, 43) = 1.68$, $p = .20$; $g = 0.05$ per each unit increase in allegiance score, 95% CI [-0.02, 0.12], $t(43) = 1.50$, $p = .14$; $g = -0.01$ per each additional year of publication, 95% CI [-0.02, 0.01], $t(43) = -0.92$, $p = .36$. However, when compared to CBT vs. EMDR comparisons, the effect of CBT vs.

integrative therapies was significantly larger when the other covariates were held constant by g = 0.45, 95% CI [0.06, 0.84], $t(43) = .03$.

Lastly, CBT-CBT comparisons, using absolute effect sizes, were included in the meta-regression to determine the extent to which effect sizes differed when comparing one type of CBT to another vs. comparing a CBT to another therapy family. CBT vs. EMDR comparisons were initially not included as there were insufficient comparisons, and disorder categories with fewer than 10 comparisons (viz., eating, neurodevelopmental, personality, sleep, and somatic disorders) were collapsed. When disorder, risk of bias, and allegiance were held constant, comparator did not significantly predict the effect size independently, $F(4, 130) = 0.50$, $p = .74$. No comparator was an individually significant predictor of effect size. The test of the model with all moderators was not significant, $F(11, 130) = 0.95$, $p = .50$, and the value of $R^2 = .06$, suggesting that these moderators explain 6% of the true variance in effects. The goodness-of-fit test indicates that there was significant unexplained variance, $\tau^2 = 0.05$, $Q(130) = 576.39$, $p < .001$, $I^2 = 77.45\%$. In all, these results suggest substantial true variance is unexplained by the included moderators.

As before, including EMDR comparisons changed the pattern of results: when controlling for the other moderators, comparator was still not a significant predictor of effect size overall, $F(5, 134) = 2.04$, $p = .08$, but EMDR-CBT comparisons emerged as an individually significant predictor of effect size: relative to comparisons with other CBTs, CBT had a lower effect size in comparisons with EMDR by g = -0.46, 95% CI [-0.79, -0.13], $t(134) = -2.78$, $p = .01$. However, this appeared to be an artifact of the arbitrary choice of positive sign for the absolute effect size of CBT-CBT comparisons when the sign of CBT-EMDR comparisons were negative. When the sign of all CBT-CBT comparisons was set to negative instead of positive, CBT-CBT and

EMDR-CBT comparisons did not differ significantly, $g = 0.21$, 95% CI [-0.13, 0.55], $t(135) = 1.23$, $p = .22$. The overall model remained non-significant, $F(12, 134) = 1.48$, $p = .14$, $R^2 = .16$, with substantial unexplained variance, $Q(134) = 597.45$, $p < .001$.

Mixed-effect subgroup analysis is shown in Table 8 below to illustrate the how effect sizes varied as a function of the comparator to CBT when not controlling for other potential moderators. Pooled within-group estimates of the variance were assumed, and the absolute value of the effect size was used for CBT-CBT comparisons. The effect sizes in Table 8 are significantly different from one another, $Q(5) = 11.58$, p = .02, implying that the difference between CBT at least one other therapy family is not precisely zero. However, no effect size meaningfully exceeded the magnitude of effect size of one type of CBT compared to another (compare: 57.8% probability of superiority for one CBT vs. another, vs. 61.7% probability of superiority of CBT vs. integrative therapy, a marginal improvement of only 3.9%). Moreover, the combined effect of all comparisons between CBT and another therapy family is lower than the overall effect when including CBT vs. CBT comparisons. These results are inconsistent with the notion that differences between therapy families exceed differences within therapy families.

Table 8

Effect Sizes and Heterogeneity Measures at Termination for Primary Disorder Measures, Relative to CBT, By Comparator

	k	*g*	95% CI	CLES	*Z*	*p*	95% PI	T
CBT	95	0.28	0.22–0.34	57.8%	9.06	< .001	-0.21–0.77	0.25
EHT	11	0.19	0.01–0.38	55.3%	--	--	-0.33–0.72	0.25
EMDR	5	-0.16	-0.41–0.09	45.5%	--	--	-0.71–0.39	0.25
Integr.	11	0.42	0.24–0.59	61.7%	4.59	<.001	-0.10–0.94	0.25
IPT	14	0.14	-0.02–.29	53.9%	--	--	-0.38–0.65	0.25
PDT	12	.20	0.03–0.37	55.6%	--	--	-0.32–0.71	0.25
Combined	50	0.19	0.09–0.28	55.3%	3.91	< .001	-0.45–0.82	0.31
Overall	139	0.24	0.19–0.29	56.7%	10.00	< .001	-0.25–0.73	0.24

Note. k = number of studies; *g* = Hedges' *g*; CI = confidence interval; CLES = Common Language Effect Size or probability of superiority; *τ* = estimate of standard deviation of true effects in *g* units; PI = prediction interval in *g* units. Positive effect sizes favour CBT. Combined = all comparators except CBT. Overall = all comparators. Estimates of *g*, *Z*, CI and PI for CBT were assigned positive signs, but all signs can be reversed arbitrarily.

Summary. For the meta-analyses of effect sizes when comparing various therapy families to CBT (i.e., the effect sizes when classifying by treatments), at termination, the main findings after accounting for outliers include the following:

1. The overall effects (where positive effect sizes favour CBT) were:
    a. $g = 0.19$, 95% CI [0.09, 0.29] for primary disorder measures, $\boldsymbol{\tau} = .31$, 95% prediction interval [-0.45, 0.83]
    b. $g = 0.09$, 95% CI [-0.08, 0.25] for secondary disorder measures, $\boldsymbol{\tau} = .35$, 95% prediction interval [-0.66, 0.83]

c. g = 0.05, 95% CI [-0.06, 0.16] for global measures, $\boldsymbol{\tau}$ = .24, 95% prediction interval [-0.45, 0.55]

2. For primary disorder outcomes, the meta-regression model containing all moderators was not significant, p = .22, and explained 14% of the variance. However, when including underpowered covariates in sensitivity analyses, some variations on the model were significant, and some significant moderators emerged.

    a. In the subgroup analysis, not controlling for other moderators, CBT had significantly better outcomes than Integrative therapies, g = 0.42, 95% CI [0.24–0.59]. However, the absolute value effect size for one CBT intervention relative to another, g = 0.28, 95% CI [0.22, 0.34], was also significant and was not meaningfully different from the effect of CBT vs. Integrative interventions (at best, <4% marginal increase in the probability of superiority).

Follow-Up

The average effect of CBT vs. other therapy families at follow-up, along with measures of heterogeneity, are presented in Table 9. The 95% CIs of all effects overlapped with the 95% PI of the summary effect, and so no outliers were removed. See Figures K5–K7 in Appendix K for forest plots.

Table 9

Effect Sizes and Heterogeneity Measures, Between-Family Effect Sizes at Follow-Up

Primary disorder measures, *k* = 26

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.17	0.03–0.30	54.8%	--	--

Q	*df*	*p*	I^2	*τ*	95% PI
182.11	25	0.001	86.27%	0.29	-0.46– 0.79

Secondary disorder measures, *k* = 14

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.07	-0.17– 0.31	52.0%	--	--

Q	*df*	*p*	I^2	*τ*	95% PI
91.37	13	< .001	85.77	.41	-0.86–1.00

Global measures, *k* = 15

Hedges' *g*	95% CI	CLES	*Z*	*p*
0.05	-0.10–.19	51.4%	--	--

Q	*df*	*p*	I^2	*τ*	95% PI
--	--	--	66.02	.21	-0.43–0.52

Note. Effect sizes of CBT vs. other therapy families; positive effect sizes favour CBT. *k* = number of studies; CI = confidence interval; CLES = Common Language Effect Size or probability of superiority; I^2 = proportion of overall variance representing variance in true effects; *τ* = estimate of standard deviation of true effects in *g* units; PI = prediction interval in *g* units.

At follow-up, the 95% confidence range for the average effect of CBT vs. other therapy families ranged from *g* = -.17 at the lowest to *g* = 0.31 at the highest (both for secondary symptom measures). A psychotherapy client could expect that by enrolling in CBT, on average, their chance of better outcomes than by enrolling in an alternative could be as low as 4.8% worse than chance and as high as an 8.7% improvement over chance. Only the effect for primary

disorder measures was significant (α = .05); the 95% confidence intervals of the average effects for secondary disorder and global measures both included zero. All effects had a large quantity of heterogeneity and moderate to large values of I^2, indicating that the average effect sizes cannot be interpreted straightforwardly: in 95% of relevant populations, the effect will favour CBT in some circumstances (by as much as g = 1.00) and the alternative therapy in others (by as much as $g = 0.86$). The Q tests were significant for primary and secondary disorder measures, but underpowered for global measures (and hence not reported here).

Although the heterogeneity measures would warrant further moderator analysis (i.e., large quantities of heterogeneity and medium-to-large values of I^2), there were an insufficient number of effects for any outcome to allow for an adequately powered meta-regression using the planned moderators. Consequently, the meta-regression results are not reported here.

Summary. For the meta-analyses of effect sizes when comparing various therapy families to CBT (i.e., the effect sizes when classifying by treatments), at follow-up, the main findings after accounting for outliers include the following:

1. The overall effects (where positive effect sizes favour CBT) were
    a. g = 0.17, 95% CI [0.03, 0.30] for primary disorder measures, $\boldsymbol{\tau}$ = .29, 95% prediction interval [-0.46, 0.79]
    b. g = 0.07, 95% CI [-0.17, 0.31] for secondary disorder measures, $\boldsymbol{\tau}$ = .41, 95% prediction interval [-0.86, 1.00]
    c. g = 0.05, 95% CI [-0.10, 0.19] for global measures, $\boldsymbol{\tau}$ = .21, 95% prediction interval [-0.43, 0.52].

Discussion

The main aims of this paper were to (1) determine whether there are differences in relative efficacy between bona fide psychotherapies (without classifying into families); (2) estimate the differences in efficacy between bona fide psychotherapy families (i.e., by classifying treatments into families and using one therapy family as a reference class), determining whether the differences between therapy families meaningfully exceed that of the differences within a family, and (3) assess the extent to which the meta-analytic results are sensitive to various conceptions of the DBV and corresponding methodological choices (e.g., standard meta-analytic effect sizes vs. absolute effect sizes). Speaking to the first and third aims, the first meta-analysis, replicating the methods of Wampold et al. (1997) (i.e., using homogeneity tests to appraise randomly-signed effects of unclassified psychotherapy comparisons), there was substantial heterogeneity, consistent with the notion that there are significant differences between psychotherapy treatments. When examining the spread of the effects, 68% of the effects fell between $g = \pm 0.29$ about zero. In the second meta-analysis, which estimated the difference between unclassified psychotherapy treatments using absolute effect sizes, the effects were similar to this approximately $g = 0.29$ value (e.g., collapsing across timepoints and outcome types, $g = 0.25$, 95% CI [0.22, 0.29]; at termination, $g = 0.27$, 0.27, and 0.30 for primary disorder, secondary disorder, and global outcomes, respectively; $g = 0.24$ for primary disorder outcomes at follow-up). These absolute effects meet the "minimally important difference" threshold of $g = 0.24$ for practical significance at the level of patient's subjective phenomenology. The effect for secondary symptom and global measures at follow-up were statistically significant, and may be practically significant at scale, but did not meet the MID threshold.

In the third meta-analysis, based on standard effect sizes and using standard random-effects meta-analytic methods, treatments were classified into therapy families, and the classes were compared using CBT as a "control." Speaking to the second and third aims, the effects of CBT vs. other treatments—although potentially meaningful at scale—generally did not meet the threshold for a phenomenologically significant effect of $g = 0.24$. For example, collapsing across timepoints and outcomes, the summary effect was $g = 0.15$, 95% CI [0.06, 0.24] in favour of CBT; for primary measures, the effect in favour of CBT was $g = 0.19$, 95% CI [0.09, 0.29] at termination and $g = 0.17$, 95% CI [0.03, 0.30] at follow-up. The effects of CBT vs. other treatment were not statistically significant for secondary symptom or for global outcome measures, either at the termination or follow-up timepoints. These effects were substantially heterogeneous. Importantly, the variance was not significantly explained by therapy family, and the effect size within a therapy family (i.e.., CBT-CBT comparisons) did not meaningfully exceed effect sizes between families. That is, when moving from a CBT-CBT comparison to a CBT-other comparison, in no case would a randomly selected client's likelihood of being in a superior treatment marginally increase by 5% or more. To the contrary, the CBT-CBT effect was *larger* than the between-family effect size in several analyses. In line with these findings, the absolute-value meta-analyses also revealed that whether a comparison was within-families or between-families did not significantly moderate the effect size.

The results of these analyses suggest several interesting and arguably surprising conclusions. Among these is the fact that by and large, none of the proposed moderators were found to have strong or consistent relations with variance in the data, despite the initial hypotheses to the contrary. This is particularly surprising given past findings that researcher allegiance has a substantial impact on the effect size between different psychotherapies (e.g.,

Luborsky et al., 1999; Munder et al., 2012; Robinson et al., 1990; Tolin et al., 2010). It is possible that the more recent studies included in this meta-analysis featured methodological improvements that weakened the relationship between allegiance and effect size that was seen in earlier meta-analyses. However, contrary to this hypothesis, the correlation between year and allegiance score was weak and non-significant, $r(154) = 0.15$, $p = .07$. Accordingly, the reason for the small impact of allegiance in this meta-analysis remains unclear. Another of the proposed moderators, methodological bias, has been less consistent in predicting variation in effect size (e.g., Goldberg et al., 2018; Tolin et al., 2010), and so it is perhaps less surprising that no strong relationship was found with variation in effect size between therapies. However, it is possible that the lack of findings in the present analysis is the result of a floor effect. That is: it is possible that the reason for a lack of relation between the studies' risk of bias and their effect size is the criteria for "high for risk of bias" (i.e., a score of 0) was too stringent for most studies to surpass. If true, then as a consequence, then all studies would simply score low on the measure, and there would be insufficient variation in the RoB2 scores to have a strong relationship with the effect size; indeed, there is evidence that in this study, variability of this moderator was low (see "Limitations" below for further elaboration of this point). Disorder, as a moderator, seemed like it may play a role in some circumstances (e.g., Tolin et al., 2010), but in general, the effect of this moderator, like the others, was small and inconsistent.

The moderator with the most direct conceptual connection to evaluating the DBV is therapy family. The precise implications of these findings for the DBV deserve some analysis. In particular, it is worth returning here to the distinctions between therapeutic techniques, treatments, and families. Whereas a technique might be defined as a specific activity undertaken in the context of psychotherapy (e.g., reappraising thoughts, exploring transference), and

whereas a treatment might be defined as a specific combination or package of techniques (often what is directly subjected to comparison in RCTs; e.g., the Mastery of your Anxiety and Worry CBT protocol for generalized anxiety disorder, Zinbarg et al., 2006; Affect Phobia therapy, McCullough, 2003), therapy families are the overarching categories of treatments united by shared theoretical explanations that guide their choice of techniques, their relational stance, and their understanding of psychopathology. Note that therapy families, though more abstract than techniques or treatment packages, are distinct from and not as abstract as the theoretical frameworks themselves.

The results of these analyses seem to suggest different interpretations for different levels. Specifically, at the level of treatments, the DBV was *not* supported: there was both a statistically and practically significant difference in absolute effect size between treatments, assuming a threshold of $g = 0.24$ for practical significance. However, this effect was heterogeneous, and moreover, no variance was attributable to whether the comparison was within a therapy family or between different therapy families. Indeed, as an example of this, comparing one type of CBT to another, on average, yielded comparable effect size differences ($g = 0.24$) as comparing a CBT intervention to a non-CBT intervention ($g = .15$). This suggests that the evidence from this study did not contradict the DBV at the level of therapy family. In short: this analysis found differences between treatments, but these do not appear to be attributable to differences between therapy families.

Several alternatives may exist for explaining the differences between treatments, given that therapy families do not appear to do so. One option is to consider if a lower level of abstraction—that is, explain the differences between treatments in terms of differences in their therapeutic techniques. This initially seems implausible, as therapy families were in this study

defined in terms of their characteristic techniques using the MULTI: certain techniques were treated as proprietary to certain families. Nevertheless, the techniques defining each family were diverse within that family; it may well be that some techniques are more efficacious than others, and that there was sufficient diversity of techniques within each family that efficacious techniques were found evenly among therapy families, but not among treatments. This could explain how therapy families could explain very little variation in the effects, even as there appeared to be more effective treatments than others. Alternatively, in a more deflationary account, the effect size differences could have less to do with technique efficacy and more to do with technique similarity. If some treatments have highly similar techniques to one another, they will show smaller effect sizes when compared against each other than those with highly dissimilar techniques. If treatments with similar and dissimilar techniques have an even distribution among the families, then effect size will be correlated with treatment but not family.[6]

To elaborate how such an explanation could work, consider a hypothetical example: within the CBT umbrella, one might imagine that there would be smaller within-family effect size differences when comparing one exposure treatment to another (either due to being similarly efficacious, or due to technical similarity); one might also imagine larger within-family effect sizes when comparing exposure-based treatment to one based on problem-solving and relaxation (if not because of differential efficacy, than at least because they are apparently technically dissimilar). Likewise, for an EHT-EHT comparison, one might imagine smaller differences between two treatments based on similar techniques (e.g., Rogerian reflection and empathizing), and larger ones with dissimilar techniques (e.g., one based on reflection vs. one based on affect experiencing, such as chair work). Analogously, when comparing between families, one might

[6] On such an account, the notion of "therapy family" would arguably lack utility when attempting to divide up various treatments, as variability within families may exceed that within families

imagine that if certain treatments are apparently similar (e.g., both chair work and exposure involve direct processing of avoided emotions), they might have smaller differences than between-family treatments with more apparently dissimilar techniques (e.g., exposure vs. reflection, or chair work vs. relaxation). Consequently, if a meta-analysis aggregates between-family comparisons that are made based on similar treatments, but within-family comparisons based on dissimilar treatments, one might find small differences between families even while comparisons within families reveal differences between treatments. The differences between techniques could permit differences between treatments, without simultaneously necessitating differences between families.

Recalling the distinction drawn earlier (p. 46) between "common factors" (transtheoretical change mechanisms that are *unrelated* to specific therapeutic procedures) and "therapeutic processes" (transtheoretical change mechanisms that *are related* to therapeutic procedures), the results of these meta-analyses could possibly be explained at a higher level of abstraction than technique or therapy family—that is, the cross-cutting therapeutic processes. On this view, these underlying therapeutic processes, which are not proprietary to any one therapy family, are what is responsible for differences in outcome. For example, both traumatic reprocessing (EMDR) and chair work (EHT) are procedures that involve experiencing avoided emotions in a safe context; both cognitive restructuring (CBT) and exploration of transference-countertransference (PDT) involve testing expectations against evidence. (Whether "experiencing avoided emotions in a safe context" or "reality testing" are in fact among the underlying therapy processes is not what is being asserted; these are just illustrative of the basic idea.) However, even though these processes are not associated with any one family, different techniques may exploit these differences to different degrees. Moreover, even if techniques were

entirely proprietary to specific families, the extent to which any given technique effectively exploits the underlying mechanisms would be *un*related to therapy family. This even distribution of effective techniques among families might explain how treatments could differ, even if therapy family did not explain the variation.

Although this hypothesis could explain much of the data from these meta-analyses, one comparison—that between CBTs and Integrative therapies—still seemed to suggest that one therapy family (Integrative therapy) had a smaller effect size than CBT ($g = 0.42$ for primary measures at termination, vs. $g = 0.24$ for the magnitude of CBT-CBT comparisons). The Integrative therapies in these analyses were treatments that crossed the threshold for having techniques from multiple therapy families without any one family predominating. Some examples of therapies which were classified as Integrative included treatments labelled by the researchers as CBASP, schema therapy, short-term psychodynamic supportive-expressive therapy, cognitive analytical therapy, motivational enhancement therapy, and some supportive therapies (incorporating both humanistic and problem-solving behavioural techniques). It is difficult to imagine what systematic reason there could be that would cause the Integrative treatments to consistently select techniques that are less effective at accessing the underlying therapeutic processes than non-Integrative therapies. This finding may make therapeutic processes a less plausible explanation for the results. However, it is worth noting that the Integrative family included a mix of treatments that were explicitly designed to integrate two therapy families (e.g., cognitive analytic therapy) and treatments that were designed without integration in mind (e.g., supportive-expressive therapy; intended to be a Psychodynamic therapy but employing many techniques characteristic of Humanistic therapy). It is possible that the latter type of Integrative therapies may inadvertently be incorporating more generic techniques that are

less directed toward key underlying therapy change processes, whereas the former treatments may be more intentional in their reliance on key, cross-cutting processes of change. This hypothesis, however, is speculative.

One further explanation—again relying on a higher level of abstraction than therapy family—involves the non-technical common factors of Wampold and likeminded theorists. On this account, the differences seen between treatments would be explained in terms of the extent to which they exploit common factors of change such as empathy and alliance. As an example of how this model could explain the results of the analysis: a given treatment (e.g., a CBT intervention) might rely on common factors such as collaboration between therapist and patient, consensus-building on goals of therapy and the tasks with which to address them, and delivers a strong rationale that builds patients' expectancy for improvement (again, whether collaboration, alliance, and expectancy are in fact among the underlying common factors is not what is being asserted; these are just illustrative of the basic idea). When compared against another CBT that does not exploit these common factors—such as one in which therapists are enjoined to be highly prescriptive, waste no time on alliance-building, and provide little rationale—the former would be expected to have a large effect size in its favour. However, the former CBT would also be expected to have a large effect size when compared against a PDT that, say, instructed therapist to be "blank screens" at the expense of empathy and alliance. Assuming, as before, that variation in how well treatments make use of the common factors is distributed evenly between therapy families, this mechanism could explain how treatments could have differential relative efficacy without therapy family predicting the differences in effect size. The common factors explanation also appears capable of accommodating the "anomaly" of the lower effect size of Integrative therapies against CBT. Namely: if a clear rationale is a common factor, or if it

facilitates other common factors such as agreement on therapeutic tasks, persuasion, or collaboration, then it is plausible that Integrative treatments might systematically supply less cohesive theoretical rationales than treatments derived from a single theoretical family. Consequently, they may be less effective at exploiting common factors, leading to smaller effect sizes relative to CBT. (On this view, one would also hypothesize smaller effect sizes relative to any other therapy family.) Again, however, it is worth re-emphasizing that this hypothesis, as with the other proposed explanations, is currently speculative, and further empirical research would be needed to appraise them.

The conclusions drawn from these results must be considered in view of their methodological strengths and limitations. In terms of strengths, this meta-analysis of relative psychotherapy efficacy comprehensively breaks down the *overall* comparison of psychotherapies into specific disorder-based populations. This may be particularly relevant given that most meta-analyses of relative efficacy—both when looking across disorders, and when choosing a specific disorder—have been primarily based on the literature on major depressive disorder. This decision is understandable, given that most primary studies are also based on the depression literature. However, as shown in this study (i.e., the Absolute Value meta-analysis), there is some evidence that the differences between treatments is especially small for depressive disorders, particularly when compared to anxiety and trauma-related disorders. This may have led to a particular bias in how judgements have been rendered on the DBV, for which this study may hopefully offer some correction. In addition to disorder, one unique feature of this study is the disaggregation of overall results by timepoint and outcome, permitting the impact of these different "DBVs" to become apparent. This study also avails itself of stringent, conservative standards for its exclusion criteria (e.g., RCTs; no studies included that did not explicitly state

their method of diagnosis), resulting in a test with a clear and specifically formulated conception of the DBV.

There are several important limitations to this study. First, to locate all clinical trials that compared two different psychotherapies, automated searches proved to be inadequate, and so hand-searching select journals was the only available method. As a consequence, the included studies were by and large restricted to a limited selection of journals, potentially biasing the results. These journals were selected due to their relatively rigorous standards and due to being relatively prolific publishers of psychotherapy RCTs. Nevertheless, caution must be taken when interpreting these data due to the potential for selection bias.

Next, although a second scorer was available to ensure reliability of both title/abstract and full text exclusion phase, no additional rater was available for the coding of moderators due to practical limitations, potentially opening the results to be influenced unduly by personal judgement.

As data collection concluded by 2021, recent studies (2021-2026) were not included in this analysis. This raises potential concerns about limitations of the dataset (e.g., excluding more recent studies that may have smaller effect studies for CBT; Johnsen & Friborg, 2015). However, analyses in Appendix M suggest that (a) the relationship between effect size and year of publication is small and (b) a very large number of missing studies with negligible effect sizes would be needed to meaningfully change the conclusions of this study.

One methodological consideration for all meta-analyses is the GIGO (Garbage-In-Garbage-Out) principle: the conclusions of the meta-analytic study are only as good as the primary studies it summarizes (e.g., Borenstein et al., 2019). One significant limitation is that, by and large, the studies included in this analysis were rated to have a high risk of bias. 88% were

rated with a score of 0 (high risk of bias in at least one domain), and 11% rated with a score of 1 (moderate risk of bias in at least one domain), with less than 2% scoring a 2 (low risk of bias). Even when liberalizing the criteria, such that the over lowest-scoring domain was discounted when determining the overall score, 51% still scored 0 and only 11% scored low risk of bias. This overall pattern of high risk of bias in the primary studies is consistent with previous findings (e.g., Cuijpers et al., 2013; Cuijpers, Harrer, Miguel, Ciharova, & Karyotaki, 2025; see also Cuijpers et al., 2010), and it suggests that by and large, psychotherapy RCTs may not meet the same standards for quality as is common in other scientific disciplines. This would suggest that further research in this domain is sorely required.

On the other hand, it may be the case that the consistently low scores for methodological quality is due to the choice of focusing on the risk of methodological bias, as opposed to other markers of methodological quality (e.g., Foa & Meadows, 1997). Specific psychotherapy-focused rating scales for methodological quality have been criticized as arbitrary (e.g., Atkins et al., 2014) and are less widespread in the literature than the risk of bias construct. However, their specificity to psychotherapy studies arguably lends these scales greater face validity. It is possible that these measures may capture important variation not represented by measures like the RoB2—leading to the lack of differences and the possible ceiling effect seen in this study.

Due to considerations of statistical power, there were limitations as to the number of moderators that could be tested in these meta-analyses. The chosen moderators, unfortunately, did not explain much of the observed variation in the effect sizes. As a result, the interpretability of these results is limited (see earlier in the Discussion).

One of the chosen moderators was the effect of therapy family (e.g., whether studies differed in their effects when comparing between families or within them). Therapy family was

operationalized as a broad category, attempting to capture as many theoretically similar treatments as possible (e.g., construing the "CBT" category as including cognitive, behavioural, and "third-wave" interventions). However, other options were available: instead of categorizing treatments, they may have rated on a dimensional variable representing the degree to which they involve prototypical techniques from each theoretical family (for example, rating each treatment and providing a score for "CBT-ness," "dynamic-ness," etc.). Alternatively, treatments may have been categorized into more granular theoretical families (e.g., "behavioural," "cognitive," "third-wave"). It is possible that, by categorizing broadly instead of categorizing specifically or by rating dimensionally, important differences between treatments and within families were obscured. However, it is also worth noting that more fine-grained divisions would have incurred greater costs in terms of statistical power.

In order to reduce heterogeneity in the effects that were artifacts of timepoint or outcome, effects were divided into primary symptom, secondary symptom, and global outcome measures, as well as by termination and follow-up timepoints. However, it is possible that more fine-grained divisions (e.g., 6-month and 1 year follow-up) would have helped to partition out more variance in the effects, leading to cleaner interpretations of the data.

This study examined the moderating effect of researcher allegiance, given previous research which has suggested the relevance of this moderator from explaining differences in effect size between treatments (e.g., Luborsky, 1999; Munder et al., 2012). However, some studies (e.g., Falkenström et al., 2013) have found that *therapist* allegiance can also potentially impact treatment outcome. This present study is limited by omitting therapist allegiance, and future studies may wish to explore its potential role.

Many of the conclusions of this study are drawn from non-standard effect sizes—e.g., effect sizes of absolute difference between treatments. There is reason to believe these absolute effect sizes overestimate differences (e.g., Marcus et al., 2012). This would suggest that, in this study, only the meta-analyses of CBT versus others can be treated as generating straightforward summary effect sizes; caution is warranted when evaluating the "upper-limit" effect sizes that come from the meta-analyses of absolute difference.

In general, this study focused on randomized controlled trials, which provides a greater degree of internal validity and reliability. However, the trade-off is that the external validity is reduced. The results of this meta-analysis may be limited in the extent to which it accounts for effectiveness of psychotherapy in practice.

Lastly, this meta-analysis was not pre-registered, which since the outset of this research project has become an increasingly wide-spread check on methodological biases and questionable practices (e.g., so-called "p-hacking"). However, these studies were proposed to a thesis committee, which may be considered as an informal "pre-registration" of intentions and methods.

Despite these limitations, it is hoped that this meta-analysis will offer clinicians some level of guidance when making determinations about the most effective approaches to heling their patients. However, like all other meta-analyses of relative efficacy, it cannot be the *only* guide to such decisions. For example, even if two treatments are both found to effective, and both found to be equivalent in their efficacy, one treatment may have substantially more evidence for its efficacy than the other. Assuming the evidence for each is of equivalent quality, clinicians may therefore prefer that treatment not because it is superior, but because they have greater epistemic warrant for their beliefs in its efficacy.

Conclusion

The Dodo Bird Verdict—the claim that all psychotherapies are equally effective—is the sort of assertion that (I have argued) can neither be refuted nor supported by these or any meta-analytic results. What is meant by "all," "psychotherapies," "equally," and "effective" in the context of psychotherapy outcome research are contentious. Accordingly, researchers may find that different, specific claims are supported by these data to greater and lesser degrees. The claim, on the one hand, that one can never find statistically significant, stable, and practically meaningful differences between any two *bona fide* treatment packages—regardless of the population, outcome measure, or timepoint—may be imperiled by this meta-analysis. On the other hand, by and large no evidence was found to support the notion that one therapy family stands as first among equals: no consistent differences in efficacy between therapy families were detected, nor was there strong evidence that specific families have superior outcomes for particular clinical populations. Perhaps most crucially, the methodological quality of the included studies was low; indeed, they raise a concern that the primary psychotherapy outcome literature is below the standards of other health fields, jeopardizing the evidentiary weight of the results altogether. In line with other calls in the literature (e.g., Cuijpers, 2016), there is a need for the funding for high-quality, direct comparisons of psychotherapies so that firmer conclusions can be drawn to guide treatment decisions for clinicians and patients. Only when our best evidence guides our best care can we conclude that *everybody* has won, and all will have prizes.

General Discussion and Conclusion

The image of the Dodo bird, rendering judgements about its haphazard race, is a potent one. Another such image comes from William James (1907/2013, Lecture II): the image of a squirrel, corkscrewing around a tree trunk. James describes two parties squabbling about just such a squirrel, along with a person on the opposite side of the tree trying to see it. The person, attempting to see the squirrel, goes all the way around the tree, but the squirrel moves just as quickly so that it always remains on the opposite side of the trunk. The dispute that James reports is this: the person goes around the tree, but do they *go around the squirrel*? James answers as follows:

> "Which party is right," I said, "depends on what you *practically mean* by 'going round' the squirrel. If you mean passing from the north of him to the east, then to the south, then to the west, and then to the north of him again, obviously the man does go round him, for he occupies these successive positions. But if on the contrary you mean being first in front of him, then on the right of him, then behind him, then on his left, and finally in front again, it is quite as obvious that the man fails to go round him […] Make the distinction, and there is no occasion for any farther dispute. You are both right and both wrong according as you conceive the verb 'to go round' in one practical fashion or the other."

I have argued that, in the psychotherapy outcome literature, the verdict on therapeutic equivalence is less like running the Dodo's race and more like a dispute about circumnavigating a squirrel. In the methodological review, we made the case for several practical matters that may impact whether a given meta-analysis does or does not support the DBV—including whether or how one classifies treatments into superordinate families, what populations one focuses on, and

sources of potential bias. In the meta-analytic study, I replicated the meta-analytic methods of previous studies, but I also included these factors as moderators to assess their impact on the result.

In the methodological review, we made the case that there would be some decisions that would be central to any test of the DBV that aimed for methodological stringency. First, we argued for meta-analyzing direct, head-to-head comparisons between psychotherapies, rather than inferring effects indirectly (e.g., by comparison to a control treatment or by network meta-analysis), due to the threats to validity introduced by either confounding variables or inappropriate assumptions of transitivity. We also argued for using meta-analytic methods that, at the time, were only just becoming common in the literature, such as random effects modeling, outlier and publication bias analysis, prediction intervals, homogeneity tests, and (when warranted by the homogeneity tests), moderator analysis to explore sources of heterogeneity.

Additionally, we argued that there could be different "DBVs"—different conceptual approaches that lead to different meta-analytic outcomes. Specifically, we contended that there may be a separate verdict on outcome equivalency depending on whether one analyzed *bona fide* clinical populations (e.g., reliably diagnosed with mental health disorders) or not (e.g., problems with academic achievement). Similarly, we made the case that without disaggregating treatments into treatment classes (e.g., CBT, Psychodynamic, etc.), similarities between within-family treatments could drown out differences between between-family treatments. This could yield a DBV for one understanding of "psychotherapy equivalence," but not another, arguably more widespread, understanding. We advanced the importance of separate meta-analyses for different timepoints (i.e., determining whether there were separate "verdicts" for post-therapy vs. long-term follow-up) and for different measures of outcome (i.e., whether the DBV was different for

disorder-specific outcomes vs. global well-being outcomes). Accordingly, we contended that meta-analyses of the DBV should first test the verdict "overall" (i.e., without classifying), as a global test to address multiple comparisons, before classifying into therapy family and disaggregating by timepoint, and by outcome. We also advocated for testing moderators of heterogeneity of the effects, which in addition to therapy family and disorder, might minimally include sources of bias such as researcher allegiance and methodological biases.

The meta-analytic study was an attempt to put into practice as many of these recommendations as was practically feasible. In so doing, I indeed found evidence that the DBV is supported under some conceptions but not others. When the effect sizes from head-to-head comparisons of treatments were aggregated with randomly distributed signs, there was significant variation according to the test of heterogeneity, with approximately 68% of effect size estimated to fall between $g = \pm 0.29$, and 95% predicted to fall between $g = -0.55$ and $g = 0.59$. When the absolute value of the differences between treatments were meta-analyzed, across all time points and outcome measures, the mean difference was $g = 0.25$, 95% CI [0.22, 0.29]. Similarly, in the analysis that used CBT as a reference class, across timepoints and outcomes, CBT had a statistically significant superiority of effect over other treatments, $g = 0.15$ with a 95% confidence interval of [0.06, 0.24]. In the subgroup analysis, there was evidence of superiority of CBT against EHT, IPT, and integrative therapies. These findings are inconsistent with a DBV that is concerned with differences between *bona fide* interventions at the level of specific packages or treatments. However, in the absolute value analysis, across timepoints and outcome, therapy family did not significantly explain heterogeneity in the summary effect; additionally, there were no differences between therapy families when the analysis was limited to between-family comparisons. Similarly, in the subgroup analysis across timepoints and

outcomes, when different therapy families were compared to CBT as a reference class, no effect size meaningfully exceeded the magnitude of effect size of one type of CBT compared to another. Moreover, the statistically significant superiority of CBT over other treatments, g = 0.15, translates to only a 54% chance that someone selected from among those treated with CBT would have a superior outcome relative to someone picked from a those treated with a non-CBT. It is difficult to argue that this is practically meaningful, especially given that the chance of having a superior outcome for someone in one CBT vs. another was found to be 57%. These findings are in line with a DBV that is concerned with the equivalence of different therapy families or approaches. Contrary to expectation, disorder, methodological bias, and allegiance bias largely did not explain variation in any of the effects. Like any good compromise, these outcomes are likely to deeply dissatisfy partisans for any strong viewpoint on the DBV.

Additionally, in line with the "Squirrel Verdict," the meta-analyses found significant and substantial heterogeneity in all summary effects, suggesting that the effect of psychotherapy is different under different study characteristics. Unexpectedly, many of the methodological factors that I hypothesized might be critical explained little variation in the effect sizes. In the absolute value meta-analysis, disorder was an individually significant moderator of effect size, such that differences between treatments were highest for anxiety and trauma-related disorders (both cutting across outcomes and timepoints, and for primary measures at termination); researcher allegiance was also individually significant for the absolute value meta-analysis when cutting across outcomes and timepoints, and also for primary measures at both termination and follow-up. However, collectively, the models containing the moderators tended to not be significant, and they tended to explain only a minute proportion of the total variance. In the meta-analyses using

CBT as a reference class, no moderator in the meta-analyses was significant, and again, the models explained a minimal proportion of the variance.

A potential limitation when attempting to interpret this analysis may be related to its data collection, which concluded in 2021. Recent studies (2021-2026) were consequently not included in this analysis, raising potential concerns about the age of the dataset. For example, more recent studies may exploit higher-quality, state-of-the-science methods relative to older ones. Even more specifically, there are some concerns that the effect size of CBT may be diminishing over time (Johnsen & Friborg, 2015; but see Cristea et al., 2017, and Hofmann et al., 2025). If true, then a concern arises that potential excluding more recent studies could bias the results in favour of CBT. However, meta-regression analyses of the relationship between effect size and year of publication (see Appendix M) are not consistent with any such statistically significant relationship, either across all included years (1980-2021) or when examining just the most recent years (2015-2021). Moreover, none of the Cognitive Contrast meta-analyses produced effect sizes in favour of CBT exceeding the "minimal important difference" threshold of $g = 0.24$; hence, if recent CBT studies with small or negative effect sizes were missing, they would not alter the conclusions of the analyses. Lastly, an Orwin's fail-safe N analysis (see Appendix M) revealed that, in order to reduce the Cognitive Contrast analyses below the secondary threshold for effects at scale (i.e., below $g = 0.05$ or below statistical significance), between 76 studies (for the analysis of primary measures at follow-up) and 131 studies (for primary measures at termination) with an effect size of $g = 0.00$ would need to be missing. Given that only 165 relevant effects were found in four decades, only approximately *25* recent studies would be expected to be missing. Consequently, the effect of these potential studies on the conclusions of the analyses is expected to be negligible.

With much heterogeneity detected in the meta-analyses and little of it explained, one can only speculate as to the circumstances under which some therapies may show better outcomes than others in RCTs. Some such speculations were outlined in the discussion section of the meta-analytic study, including the precise technical makeup of treatment packages or the potential role of cross-cutting change mechanisms, whether primarily interpersonal in nature (i.e., "common factors"; e.g., Wampold & Imel, 2015) or related to technique and theory (i.e., "therapeutic processes"; e.g., Norcross, 2005). Other possibilities may include patient population characteristics not appraised by our meta-analysis (e.g., differences between RCTs related to participant age, experiences with socioeconomic class, sex, degree of symptom distress, experiences with racialization, gender, level of motivation, comorbidities, etc.), or therapist characteristics (e.g., differences between RCTs in terms of therapist training, adeptness, ability to form alliance, etc.).

Although these hypotheses were not tested by the meta-analyses in this dissertation, there is a large literature of process-outcome research that, over the decades, has been accumulating direct non-meta-analytic evidence about the most critical factors for psychotherapy efficacy. Reviewing such a large corpus of work is beyond the scope of this dissertation; however, the interested reader is directed to overviews such as Barkham et al. (2021), Castonguay et al. (2019a), Lambert (2013), Norcross and Lambert (2019), and Norcross and Wampold (2019), and (for a CBT-specific discussion of change mechanisms) Hayes and Hofmann (2018). This literature may well provide a more fruitful approach to understanding the observed variation in treatment outcome comparisons than meta-analyses of the DBV have provided by themselves.

The lack of forthcoming explanations for the heterogeneity in the results raise the question as to what practical conclusions might be drawn from this research. Part of the impetus

for limiting the statistical analyses to so-called *bona fide* clinical populations was so that these results would be most directly relevant to the public health domain, including clinicians' decision-making, health policy, and training of future clinicians for the healthcare context. Obviously, these data *cannot* comprehensively guide such decisions. For one, as outlined before, these data are consistent with differences between specific treatments, but give no guidance as to which treatments are superior, or why. Moreover, evidentiary warrant for treatment differences (as addressed in this dissertation) must be balanced against epistemic warrant—even if treatments are equally effective, for example, the evidence base for one treatment may substantially outstrip the other, which alone may influence one's credence in the effectiveness of a given treatment.

Nevertheless, with the above caveats in mind, these results do appear to have at least some implications for clinical care and training. First, the fact that there was evidence of differences between specific treatments suggests that therapists should, indeed, pay attention to matters of therapeutic procedure to maximize outcomes for their clients. The refrain one sometimes hears from clinicians that therapist actions don't matter—often too casually justified by reference to the DBV literature—does not appear to hold. This is not only apparent in light of the results of these meta-analyses, but also because the DBV was only ever purported to hold among *bona fide* treatments, a rarefied class that excludes all treatments that are not guided by psychological theory. For example, *bona fide* supportive treatments (i.e., those from the client-centred tradition) may be no less effective than other *bona fide* treatments, whereas non-*bona fide* supportive therapy (e.g., nondirective control treatments) may evince lower effect sizes (e.g., Cuijpers, Miguel, Ciharova, Harrer, & Karyotaki, 2024).

Simultaneously, these results supply no evidence that specific treatment families are the decisive factor differentiating more and less effective treatments. This has potentially significant implications for both public health policy and training policies. From the standpoint of differential efficacy (again, bracketing epistemic considerations), these data are not consistent with training policies that prioritize one class of treatment families over another. Rather, a more non-sectarian and arguably optimistic message is implied: many *bona fide* treatments are effective, and training mental health clinicians in any one of them can lead to positive outcomes for clients. Indeed, these data suggest that pluralistic and diverse training in *bona fide* treatments, along with the best evidence from the literature on process-outcome research, can help to expose trainees to the wide variety of factors that may be crucial to optimizing benefits for patients. Similarly, these data do not support a health care model wherein mental health funding is prioritized for particular classes of psychotherapy; rather, these data are consistent with pluralism when funding *bona fide* treatments. Indeed, research seems to suggest that a helpful predictor of superior treatment efficacy is the very quality of being an evidence-based or *bona fide* treatment—and the extent to which "treatments as usual" approximate this criterion may moderate their relative effect size (e.g., Wampold et al., 2011). In 2021, the UK's Talking Psychotherapy program (formerly Improving Access to Psychotherapy, IAPT), expanded its mandate—previously largely focused on promulgating access to CBTs—to focus on hiring therapists working in other evidence-based non-CBTs, such as IPT and dynamic-interpersonal therapy (UK Council for Psychotherapy, n.d.). Such a move would comport with the results of my meta-analytic study. Additionally, in line with the implications for training, the most efficient way for public health institutions to maximize benefit for their patients may be to simply hire graduates whose clinical training prepares them to (a) work with a diverse set of roughly

equivalently effective treatment approaches and (b) understand and apply the research connecting therapeutic process and outcome. Such training may provide clinicians with the expertise to select the particular treatments (and interpersonal approaches; e.g., Norcross & Lambert, 2019; Norcross & Wampold, 2019) that are likely to benefit particular clients—rather than relying on treatment classes that may be too coarse-grained to do the job.

Given that the bodies of research both for process-outcome studies and comparative RCTs are quite limited as they currently stand, I echo the calls in the literature (e.g., Castonguay et al., 2019b; Cuijpers, 2016) for further study in these areas. Specifically, I encourage an emphasis on higher-quality and more adequately powered RCTs, along with dismantling designs and detailed video recording of therapy processes. Simultaneously, I also recognize that the dearth of such high-quality research is due in no small part to their practical difficulty, their expensiveness, and the "scandalous" underfunding and undervaluing of psychotherapy by funding agencies (Therapy deficit, 2012). Although re-orienting priorities at the systems level can be slow work, we may perhaps adopt a measure of cautious optimism about technological change in the past years. For example, despite their well-documented shortcomings (e.g., Athaluri, 2023; Strubell et al., 2020), machine learning and "artificial intelligence" technologies appear to be extremely adept at classifying patterns (e.g., Jumper et al., 2021) and have already been applied to the automatic coding of psychotherapy video recordings (Doorn et al., 2025) and extracting data from RCTs (Kataoka et al., 2025). As these technologies proliferate, they may facilitate less complicated and less expensive studies of psychotherapy process.

As the evidence base continues to grow, it is my hope that the components of my dissertation has provided useful insight into the methodological, conceptual, and empirical issues to contend with when appraising the DBV. As the field progresses, it may well be that the best

way to answer the question of the DBV is to un-ask it, the best way to solve it is to dissolve it, and the best way to develop it is to evolve from it.

References

References marked with an asterisk indicate studies included in the meta-analysis. References marked with a dagger indicate meta-analyses reviewed in Table 1.

Aafjes-van Doorn, K., Cicconet, M., Cohn, J. F., & Aafjes, M. (2025). Predicting working alliance in psychotherapy: A multi-modal machine learning approach. *Psychotherapy Research*, *35*(2), 256–270. https://doi.org/10.1080/10503307.2024.2428702

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*(1), 129–133. https://doi.org/10.1037/0033-2909.97.1.129

*Afshari, B., & Hasani, J. (2020). Study of dialectical behavior therapy versus cognitive behavior therapy on emotion regulation and mindfulness in patients with generalized anxiety disorder. *Journal of Contemporary Psychotherapy*, *50*(4), 305–312. https://doi.org/10.1007/s10879-020-09461-9

*Agras, W. S., Schneider, J. A., Arnow, B., Raeburn, S. D., & Telch, C. F. (1989). Cognitive-behavioral treatment with and without exposure plus response prevention in the treatment of bulimia nervosa: A reply to Leitenberg and Rosen. *Journal of Consulting and Clinical Psychology.*, *57*(6), 778–779. https://doi.org/10.1037/0022-006X.57.6.778

*Agras, W. S., Walsh, B. T., Fairburn, C. G., Wilson, G. T., & Kraemer, H. C. (2000). A multicenter comparison of cognitive-behavioral therapy and interpersonal psychotherapy for bulimia nervosa. *Archives of General Psychiatry*, *57*(5), 459–466. https://doi.org/10.1001/archpsyc.57.5.459

Aguilar, M. I., & Hart, R. (2005). Oral anticoagulants for preventing stroke in patients with non-valvular atrial fibrillation and no previous history of stroke or transient ischemic

attacks. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD001927.pub2

Ahn, H.-n., & Wampold, B. E. (2001). Where oh where are the specific ingredients? A meta-analysis of component studies in counseling and psychotherapy. *Journal of Counseling Psychology*, *48*(3), 251–257. https://doi.org/10.1037/0022-0167.48.3.251

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.).

Anderson, E. M., & Lambert, M. J. (1995). Short-term dynamically oriented psychotherapy: A review and meta-analysis. *Clinical Psychology Review*, *15*(6), 503–514. https://doi.org/10.1016/0272-7358(95)00027-M

Andrews, G., & Harvey, R. (1981). Does psychotherapy benefit neurotic patients? A reanalysis of the Smith, Glass, and Miller Data. *Archives of General Psychiatry*, *38*(11), 1203–1208. https://doi.org/10.1001/archpsyc.1981.01780360019001

Anglican Communion Office. (n.d.). *The Anglican communion covenant.* Anglican Communion. https://www.anglicancommunion.org/media/99905/The_Anglican_Covenant.pdf

Ansbacher, H. L., & Ansbacher, R. R. (Eds.). (1956). *The individual psychology of Alfred Adler*. Basic Books.

*Arch, J. J., Eifert, G. H., Davies, C., Vilardaga, J. C. P., Rose, R. D., & Craske, M. G. (2012). Randomized clinical trial of cognitive behavioral therapy (CBT) versus acceptance and commitment therapy (ACT) for mixed anxiety disorders. *Journal of Consulting and Clinical Psychology*, *80*(5), 750–765. https://doi.org/10.1037/a0028310

*Arntz, A., & Van Den Hout, M. (1996). Psychological treatments of panic disorder without agoraphobia: Cognitive therapy versus applied relaxation. *Behaviour Research and Therapy*, *34*(2), 113–121. https://doi.org/10.1016/0005-7967(95)00061-5

Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Curēus*, *15*(4), Article e37432. https://doi.org/10.7759/cureus.37432

Atkins, P. W. B., Ciarrochi, J., Gaudiano, B. A., Bricker, J. B., Donald, J., Rovner, G., Smout, M., Livheim, F., Lundgren, T., & Hayes, S. C. (2017). Departing from the essential features of a high quality systematic review of psychotherapy: A response to Öst (2014) and recommendations for improvement. *Behaviour Research and Therapy*, *97*, 259–272. https://doi.org/10.1016/j.brat.2017.05.016

†Baardseth, T. P., Goldberg, S. B., Pace, B. T., Wislocki, A. P., Frost, N. D., Siddiqui, J. R., Lindemann, A. M., Kivlighan, D. M., Laska, K. M., Del Re, A. C., Minami, T., & Wampold, B. E. (2013). Cognitive-behavioral therapy versus other therapies: Redux. *Clinical Psychology Review*, *33*(3), 395–405. https://doi.org/10.1016/j.cpr.2013.01.004

*Barber, J. P., Milrod, B., Gallop, R., Solomonov, N., Rudden, M. G., McCarthy, K. S., & Chambless, D. L. (2020). Processes of therapeutic change: Results from the Cornell-Penn Study of Psychotherapies for Panic Disorder. *Journal of Counseling Psychology, 67*(2), 222–231. https://doi.org/10.1037/cou0000417

Barkham, M., & Lambert, M. (2021). The efficacy and effectiveness of psychological therapies. In Barkham, M., Lutz, W., & Castonguay, L. G. (Eds.). *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th ed., pp. 135–189). John Wiley & Sons.

Barkham, M., Lutz, W., & Castonguay, L. G. (Eds.). (2021). *Bergin and Garfield's handbook of psychotherapy and behavior change* (7th ed.). John Wiley & Sons.

Barlow, D. H. (2004). Psychological treatments. *American Psychologist, 59*(9), 869–878. https://doi.org/10.1037/0003-066X.59.9.869

*Barlow, D. H., Craske, M. G., Cerny, J. A., & Klosko, J. S. (1989). Behavioral treatment of panic disorder. *Behavior Therapy*, *20*(2), 261–282. https://doi.org/10.1016/S0005-7894(89)80073-5

*Barlow, D. H., Farchione, T. J., Bullis, J. R., Gallagher, M. W., Murray-Latin, H., Sauer-Zavala, S., Bentley, K. H., Thompson-Hollands, J., Conklin, L. R., Boswell, J. F., Ametaj, A., Carl, J. R., Boettcher, H. T., & Cassiello-Robbins, C. (2017). The Unified Protocol for transdiagnostic treatment of emotional disorders compared with diagnosis-specific protocols for anxiety disorders: A randomized clinical trial. *JAMA Psychiatry, 74*(9), 875–884. https://doi.org/10.1001/jamapsychiatry.2017.2164Barlow, Rapee, Brown, 1992

†Barnicot, K., Michael, C., Trione, E., Lang, S., Saunders, T., Sharp, M., & Crawford, M. J. (2020). Psychological interventions for acute psychiatric inpatients with schizophrenia-spectrum disorders: A systematic review and meta-analysis. *Clinical Psychology Review*, *82*, Article 101929. https://doi.org/10.1016/j.cpr.2020.101929

*Barrowclough, C., King, P., Colville, J., Russell, E., Burns, A., & Tarrier, N. (2001). A randomized trial of the effectiveness of cognitive-behavioral therapy and supportive

counseling for anxiety symptoms in older adults. *Journal of Consulting and Clinical Psychology, 69*(5), 756–762. https://doi.org/10.1037/0022-006X.69.5.756

Beck, A. T. (2007, February 21). Does cognitive therapy = cognitive behavior therapy? *CBT Insights*. https://beckinstitute.org/blog/does-cognitive-therapy-cognitive-behavior-therapy/

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression* (1st ed.). Guilford Press.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory–II*. Psychological Corporation.

*Bellino, S., Zizza, M., Rinaldi, C., & Bogetto, F. (2007). Combined therapy of major depression with concomitant borderline personality disorder: Comparison of interpersonal and cognitive psychotherapy. Canadian Journal of Psychiatry, 52(11), 718–725. https://doi.org/10.1177/070674370705201106

*Belloch, A., Cabedo, E., & Carrió, C. (2008). Cognitive versus behaviour therapy in the individual treatment of obsessive-compulsive disorder: Changes in cognitions and clinically significant outcomes at post-treatment and one-year follow-up. *Behavioural and Cognitive Psychotherapy.*, *36*(5), 521–540. https://doi.org/10.1017/S1352465808004451

†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. https://doi.org/10.1016/j.cpr.2007.10.005

Benish, S. G., Quintana, S., & Wampold, B. E. (2011). Culturally adapted psychotherapy and the legitimacy of myth: A direct-comparison meta-analysis. *Journal of Counseling Psychology*, *58*(3), 279–289. https://doi.org/10.1037/a0023626

Bentham, J. (1961). *An introduction to the principles of morals and legislation*. In *The utilitarians* (pp. 7–398). Doubleday. https://hdl.handle.net/2027/uc1.32106005889057 (Original work published 1789)

Bergin, A. E. (1971). The evaluation of therapeutic outcomes. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (1st ed., pp. 217-270). John Wiley & Sons.

*Bernecker, S. L., Constantino, M. J., Atkinson, L. R., Bagby, R. M., Ravitz, P., & McBride, C. (2016). Attachment style as a moderating influence on the efficacy of cognitive-behavioral and interpersonal psychotherapy for depression: A failure to replicate. *Psychotherapy*, *53*(1), 22–33. https://doi.org/10.1037/pst0000036

*Beutel, M. E., Scheurich, V., Knebel, A., Michal, M., Wiltink, J., Graf-Morgenstern, M., ... & Subic-Wrana, C. (2013). Implementing panic-focused psychodynamic psychotherapy into clinical practice. *The Canadian Journal of Psychiatry*, *58*(6), 326–334. https://doi.org/10.1177/070674371305800604

*Beutel, M. E., Scheurich, V., Knebel, A., Michal, M., Wiltink, J., Graf-Morgenstern, M., Tschan, R., Milrod, B., Wellek, S., & Subic-Wrana, C. (2013). Implementing panic-focused psychodynamic psychotherapy into clinical practice. *Canadian Journal of Psychiatry*, *58*(6), 326–334. https://doi.org/10.1177/070674371305800604

†Bisson, J. I., Ehlers, A., Matthews, R., Pilling, S., Richards, D., & Turner, S. (2007). Psychological treatments for chronic post-traumatic stress disorder: Systematic review

and meta-analysis. *The British Journal of Psychiatry*, *190*(2), 97–104. https://doi.org/10.1192/bjp.bp.106.021402

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*. https://doi.org/10.1002/14651858.CD003388.pub4

*Blakey, S. M., Abramowitz, J. S., Buchholz, J. L., Jessup, S. C., Jacoby, R. J., Reuman, L., & Pentel, K. Z. (2019). A randomized controlled trial of the judicious use of safety behaviors during exposure therapy. *Behaviour Research and Therapy*, *112*, 28–35. https://doi.org/10.1016/j.brat.2018.11.010

*Blanco, C., Markowitz, J. C., Hellerstein, D. J., Nezu, A. M., Wall, M., Olfson, M., Chen, Y., Levenson, J., Onishi, M., Varona, C., Okuda, M., & Hershman, D. L. (2019). A randomized trial of interpersonal psychotherapy, problem solving therapy, and supportive therapy for major depressive disorder in women with breast cancer. *Breast Cancer Research and Treatment*, *173*(2), 353–364. https://doi.org/10.1007/s10549-018-4994-5

*Bodenmann, G., Plancherel, B., Beach, S. R. H., Widmer, K., Gabriel, B., Meuwly, N., Charvoz, L., Hautzinger, M., & Schramm, E. (2008). Effects of coping-oriented couples therapy on depression: A randomized clinical trial. *Journal of Consulting and Clinical Psychology*, *76*(6), 944–954. https://doi.org/10.1037/a0013467

Borenstein, M. (2000). The shift from significance testing to effect size estimation. In A. S. Bellack & M. Hersen (Eds.), *Comprehensive clinical psychology* (Vol. 3, pp. 313–349). Pergamon.

Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them*. Biostat, Inc. https://meta-analysis-workshops.com/download/commonmistakes.pdf

Borenstein, M., Hedges, L. E., Higgins, J. P. T., & Rothstein, H. R. (2022). *Comprehensive meta-analysis* (Version 4) [Computer software]. Biostat. https://www.meta-analysis.com/

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009*). Introduction to meta-analysis* (1st ed.). John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to meta-analysis* (2nd ed.). John Wiley & Sons.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research synthesis methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Borenstein, M., Higgins, J. P. T., Hedges, L. V., & Rothstein, H. R. (2017). Basics of meta-analysis: I2 is not an absolute measure of heterogeneity. *Research Synthesis Methods*, *8*(1), 5–18. https://doi.org/10.1002/jrsm.1230

*Borkovec, T. D., & Costello, E. (1993). Efficacy of applied relaxation and cognitive-behavioral therapy in the treatment of generalized anxiety disorder. *Journal of Consulting and Clinical Psychology*, *61*(4), 611–619. https://doi.org/10.1037/0022-006X.61.4.611

*Borkovec, T. D., & Mathews, A. M. (1988). Treatment of nonphobic anxiety disorders: A comparison of nondirective, cognitive, and coping desensitization therapy. *Journal of Consulting and Clinical Psychology*, *56*(6), 877–884. https://doi.org/10.1037/0022-006X.56.6.877

*Borkovec, T. D., Newman, M. G., Pincus, A. L., & Lytle, R. (2002). A component analysis of cognitive-behavioral therapy for generalized anxiety disorder and the role of interpersonal problems. *Journal of Consulting and Clinical Psychology*, *70*(2), 288–298. https://doi.org/10.1037/0022-006X.70.2.288

*Bramoweth, A. D., Lederer, L. G., Youk, A. O., Germain, A., & Chinman, M. J. (2020). Brief behavioral treatment for insomnia vs. cognitive behavioral therapy for insomnia: Results of a randomized noninferiority clinical trial among veterans. *Behavior Therapy*, *51*(4), 535–547. https://doi.org/10.1016/j.beth.2020.02.002

†Braun, S. R., Gregor, B., & Tran, U. S. (2013). Comparing bona fide psychotherapies of depression in adults with two meta-analytical approaches. *PloS One*, *8*(6), e68135. https://doi.org/10.1371/journal.pone.0068135

*Bruijniks, S. J. E., Lemmens, L. H. J. M., Hollon, S. D., Peeters, F. P. M. L., Cuijpers, P., Arntz, A., Dingemanse, P., Willems, L., van Oppen, P., Twisk, J. W. R., van den Boogaard, M., Spijker, J., Bosmans, J., & Huibers, M. J. H. (2020). The effects of once- versus twice-weekly sessions on psychotherapy outcomes in depressed patients. *British Journal of Psychiatry*, *216*(4), 222–230. https://doi.org/10.1192/bjp.2019.265

*Bryant, R. A., Mastrodomenico, J., Felmingham, K. L., Hopwood, S., Kenny, L., Kandris, E., Cahill, C., & Creamer, M. (2008). Treatment of acute stress disorder: A randomized controlled trial. *Archives of General Psychiatry*, *65*(6), 659–667. https://doi.org/10.1001/archpsyc.65.6.659

*Bryant, R. A., Mastrodomenico, J., Hopwood, S., Kenny, L., Cahill, C., Kandris, E., & Taylor, K. (2013). Augmenting cognitive behaviour therapy for post-traumatic stress disorder with emotion tolerance training: a randomized controlled trial. *Psychological Medicine*, *43*(10), 2153–2160. https://doi.org/10.1017/S0033291713000068

*Bryant, R. A., Moulds, M. L., Guthrie, R. M., Dang, S. T., & Nixon, R. D. V. (2003). Imaginal exposure alone and imaginal exposure with cognitive restructuring in treatment of

posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, *71*(4), 706–712. https://doi.org/10.1037/0022-006X.71.4.706

*Bryant, R. A., Sackville, T., Dang, S. T., Moulds, M., & Guthrie, R. (1999). Treating acute stress disorder: An evaluation of cognitive behavior therapy and supportive counseling techniques. *The American Journal of Psychiatry*, *156*(11), 1780–1786. https://doi.org/10.1176/ajp.156.11.1780

*Buckner, J. D., Zvolensky, M. J., Ecker, A. H., Schmidt, N. B., Lewis, E. M., Paulus, D. J., Lopez-Gamundi, P., Crapanzano, K. A., & Bakhshaie, J. (2019). Integrated cognitive behavioral therapy for comorbid cannabis use and anxiety disorders: A pilot randomized controlled trial. *Behaviour Research and Therapy*, *115*, 38–45. https://doi.org/10.1016/j.brat.2018.10.014

†Budge, S. L., Moore, J. T., Del Re, A. C., Wampold, B. E., Baardseth, T. P., & Nienhuis, J. B. (2013). The effectiveness of evidence-based treatments for personality disorders when comparing treatment-as-usual and bona fide treatments. *Clinical Psychology Review*, *33*(8), 1057–1066. https://doi.org/10.1016/j.cpr.2013.08.003

*Budney, A. J., Higgins, S. T., Radonovich, K. J., & Novy, P. L. (2000). Adding voucher-based incentives to coping skills and motivational enhancement improves outcomes during treatment for marijuana dependence. *Journal of Consulting and Clinical Psychology*, *68*(6), 1051–1061. https://doi.org/10.1037/0022-006X.68.6.1051

*Budney, A. J., Moore, B. A., Rocha, H. L., & Higgins, S. T. (2006). Clinical trial of abstinence-based vouchers and cognitive-behavioral therapy for cannabis dependence. *Journal of Consulting and Clinical Psychology*, *74*(2), 307–316. https://doi.org/10.1037/0022-006X.74.2.307

*Burke, M., Drummond, L. M., & Johnston, D. W. (1997). Treatment choice for agoraphobic women: Exposure or cognitive-behaviour therapy? *British Journal of Clinical Psychology*, *36*(3), 409–420. https://doi.org/10.1111/j.2044-8260.1997.tb01248.x

*Burling, T. A., Seidner Burling, A., & Latini, D. (2001). A controlled smoking cessation trial for substance-dependent inpatients. *Journal of Consulting and Clinical Psychology*, *69*(2), 295–304. https://doi.org/10.1037/0022-006X.69.2.295

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review, 26*(1), 17–31. https://doi.org/10.1016/j.cpr.2005.07.003

*Butler, G., Fennell, M., Robson, P., & Gelder, M. (1991). Comparison of behavior therapy and cognitive behavior therapy in the treatment of generalized anxiety disorder. *Journal of Consulting and Clinical Psychology*, *59*(1), 167–175. https://doi.org/10.1037/0022-006X.59.1.167

*Butollo, W., Karl, R., König, J., & Rosner, R. (2016). A randomized controlled clinical trial of Dialogical Exposure Therapy versus Cognitive Processing Therapy for adult outpatients suffering from PTSD after Type I trauma in adulthood. *Psychotherapy and Psychosomatics*, *85*(1), 16–26. https://doi.org/10.1159/000440726

*Capezzani, L., Ostacoli, L., Cavallo, M., Carletto, S., Fernandez, I., Solomon, R., Pagani, M., & Cantelmi, T. (2013). EMDR and CBT for cancer patients: Comparative study of effects on PTSD, anxiety, and depression. *Journal of EMDR Practice and Research*, *7*(3), 134–143. http://dx.doi.org/10.1891/1933-3196.7.3.134

Carroll, L. (2008). *Alice's adventures in wonderland*. Project Gutenberg. https://www.gutenberg.org/files/11/11-h/11-h.htm (Original work published 1865)

*Carter, J. D., Crowe, M., Carlyle, D., Frampton, C. M., Jordan, J., McIntosh, V. V. W., O'Toole, V. M., Whitehead, L., & Joyce, P. R. (2012). Patient change processes in psychotherapy: Development of a new scale. *Psychotherapy Research*, *22*(1), 115–126. https://doi.org/10.1080/10503307.2011.631195

*Carter, J. D., McIntosh, V. V., Jordan, J., Porter, R. J., Frampton, C. M., & Joyce, P. R. (2013). Psychotherapy for depression: A randomized clinical trial comparing schema therapy and cognitive behavior therapy. *Journal of Affective Disorders*, *151*(2), 500–505. https://doi.org/10.1016/j.jad.2013.06.034

†Caselli, I., Ielmini, M., Bellini, A., Zizolfi, D., & Callegari, C. (2023). Efficacy of short-term psychodynamic psychotherapy (STPP) in depressive disorders: A systematic review and meta-analysis. *Journal of Affective Disorders*, *325*, 169–176. https://doi.org/10.1016/j.jad.2022.12.161

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., & Moffitt, T. E. (2014). The *p* factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Castonguay, L. G., Constantino, M. J., & Beutler, L. E. (2019a). Implementing evidence-based principles of therapeutic change: A bidirectional collaboration between clinicians and researchers. In L. G. Castonguay, M. J. Constantino, & L. E. Beutler (Eds.), *Principles of change: how psychotherapists implement research in practice* (pp. 3–12). Oxford University Press.

Castonguay, L. G., Constantino, M. J., & Beutler, L. E. (Eds.). (2019b). *Principles of change: how psychotherapists implement research in practice*. Oxford University Press.

†Chen, L., Zhang, G., Hu, M., & Liang, X. (2015). Eye movement desensitization and reprocessing versus cognitive-behavioral therapy for adult posttraumatic stress disorder: Systematic review and meta-analysis. *The Journal of Nervous and Mental Disease*, *203*(6), 443–451. https://doi.org/10.1097/NMD.0000000000000306

Cipriani, A., Higgins, J. P. T., Geddes, J. R., & Salanti, G. (2013). Conceptual and technical challenges in network meta-analysis. *Annals of Internal Medicine*, *159*(2), 130–137. https://doi.org/10.7326/0003-4819-159-2-201307160-00008

*Clark, D. M., Ehlers, A., Hackmann, A., McManus, F., Fennell, M., Grey, N., Waddington, L., & Wild, J. (2006). Cognitive therapy versus exposure and applied relaxation in social phobia: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *74*(3), 568–578. https://doi.org/10.1037/0022-006X.74.3.568

*Clark, D. M., Salkovskis, P. M., Hackmann, A., Middleton, H., Anastasiades, P., & Gelder, M. (1994). A comparison of cognitive therapy, applied relaxation and imipramine in the treatment of panic disorder. *British Journal of Psychiatry*, *164*(6), 759–769. https://doi.org/10.1192/bjp.164.6.759

*Clark, D., Salkovskis, P., Hackmann, A., Wells, A., Fennell, M., Ludgate, J., Ahmad, S., Richards, H., & Gelder, M. (1998). Two psychological treatments for hypochondriasis. A randomised controlled trial. *British Journal of Psychiatry*, *173*(3), 218–225. https://doi.org/10.1192/bjp.173.3.218

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, *37*(3/4), 256–266.

Cochran, W. G., & Carroll, S. P. (1953). A sampling investigation of the efficiency of weighting inversely as the estimated variance. *Biometrics*, *9*(4), 447–459. https://doi.org/10.2307/3001436

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Earlbaum Associates.

*Connolly Gibbons, M. B., Gallop, R., Thompson, D., Luther, D., Crits-Christoph, K., Jacobs, J., Yin, S., & Crits-Christoph, P. (2016). Comparative effectiveness of cognitive therapy and dynamic psychotherapy for major depressive disorder in a community mental health setting: A randomized clinical noninferiority trial. *JAMA Psychiatry*, *73*(9), 904–912. https://doi.org/10.1001/jamapsychiatry.2016.1720

*Constantino, M. J., Marnell, M. E., Haile, A. J., Kanther-Sista, S. N., Wolman, K., Zappert, L., & Arnow, B. A. (2008). Integrative cognitive therapy for depression: A randomized pilot comparison. *Psychotherapy*, *45*(2), 122–134. https://doi.org/10.1037/0033-3204.45.2.122

*Constantino, Marnell, Haile, Kanther-Sista, Wolman, Zappert, Arnow, 2008

*Cottraux, J., Note, I., Yao, S. N., de Mey-Guillard, C., Bonasse, F., Djamoussian, D., Mollard, E., Note, B., & Chen, Y. (2008). Randomized controlled comparison of cognitive behavior therapy with rogerian supportive therapy in chronic post-traumatic stress disorder: A 2-year follow-up. *Psychotherapy and Psychosomatics*, *77*(2), 101–110. https://doi.org/10.1159/000112887

*Cottraux, J., Note, I., Yao, S. N., Lafont, S., Note, B., Mollard, E., Bouvard, M., Sauteraud, A., Bourgeois, M., & Dartigues, J.-F. (2001). A randomized controlled trial of cognitive therapy versus intensive behavior therapy in obsessive compulsive

disorder. *Psychotherapy and Psychosomatics*, *70*(6), 288–297. https://doi.org/10.1159/000056269

*Craske, M. G., Meuret, A. E., Ritz, T., Treanor, M., Dour, H., & Rosenfield, D. (2019). Positive affect treatment for depression and anxiety: A randomized clinical trial for a core feature of anhedonia. *Journal of Consulting and Clinical Psychology*, *87*(5), 457–471. https://doi.org/10.1037/ccp0000396

Creed, T. A., Wolk, C. B., Feinberg, B., Evans, A. C., & Beck, A. T. (2016). Beyond the label: Relationship between community therapists' self-report of a cognitive behavioral therapy orientation and observed skills. *Administration and Policy in Mental Health and Mental Health Services Research*, *43*(1), 36–43. https://doi.org/10.1007/s10488-014-0618-5

†Cristea, I. A., Huibers, M. J. H., David, D., Hollon, S. D., Andersson, G., & Cuijpers, P. (2015). The effects of cognitive behavior therapy for adult depression on dysfunctional thinking: A meta-analysis. *Clinical Psychology Review*, *42*, 62–71. https://doi.org/10.1016/j.cpr.2015.08.003

Cristea, I. A., Stefan, S., Karyotaki, E., David, D., Hollon, S. D., & Cuijpers, P. (2017). The effects of cognitive behavioral therapy are not systematically falling: A revision of Johnsen and Friborg (2015). *Psychological Bulletin*, *143*(3), 326–340. https://doi.org/10.1037/bul0000062

Crits-Christoph, P. (1997). Limitations of the dodo bird verdict and the role of clinical trials in psychotherapy research: Comment on Wampold et al. (1997). *Psychological Bulletin, 122*(3), 216–220. https://doi.org/10.1037/0033-2909.122.3.216

*Crits-Christoph, P., Siqueland, L., Blaine, J., Frank, A., Luborsky, L., Onken, L. S., Muenz, L. R., Thase, M. E., Weiss, R. D., Gastfriend, D. R., Woody, G. E., Barber, J. P., Butler, S.

F., Daley, D., Salloum, I., Bishop, S., Najavits, L. M., Lis, J., Mercer, D., … Beck, A. T. (1999). Psychosocial treatments for cocaine dependence: National Institute on Drug Abuse Collaborative Cocaine Treatment Study. *Archives of General Psychiatry*, *56*(6), 493–502. https://doi.org/10.1001/archpsyc.56.6.493

Cronbach, L.J. (1982). Prudent aspirations for social inquiry (pp. 61–81). In Kruskal, W.H. (Ed.), *The social sciences: Their nature and uses*. The University of Chicago Press.

Cuijpers, P. (2016). Are all psychotherapies equally effective in the treatment of adult depression? The lack of statistical power of comparative outcome studies. *Evidence Based Mental Health*, *19*(2), 39–42. https://doi.org/10.1136/eb-2016-102341

Cuijpers, P. (2025, September 14). *Recent publications*. https://www.pimcuijpers.com/blog/recent-publications/

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Canadian Journal of Psychiatry*, *58*(7), 376–385. https://doi.org/10.1177/070674371305800702

Cuijpers, P., Driessen, E., Hollon, S. D., van Oppen, P., Barth, J., & Andersson, G. (2012). The efficacy of non-directive supportive therapy for adult depression: A meta-analysis. *Clinical Psychology Review*, *32*(4), 280–291. https://doi.org/10.1016/j.cpr.2012.01.003

Cuijpers, P., Harrer, M., Miguel, C., Ciharova, M., & Karyotaki, E. (2025). Five decades of research on psychological treatments of depression: A historical and meta-analytic overview. *The American Psychologist*, *80*(3), 297–310. https://doi.org/10.1037/amp0001250

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., … 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., Pot, A. M., Park, M., & Reynolds, C. F. (2014). Managing depression in older age: Psychological interventions. *Maturitas*, *79*(2), 160–169. https://doi.org/10.1016/j.maturitas.2014.05.027

Cuijpers, P., Karyotaki, E., Reijnders, M., & Ebert, D. D. (2019). Was Eysenck right after all? A reassessment of the effects of psychotherapy for adult depression. *Epidemiology and Psychiatric Sciences*, *28*(1), 21–30. https://doi.org/10.1017/S2045796018000057

Cuijpers, P., Miguel, C., Ciharova, M., Harrer, M., & Karyotaki, E. (2024). Non-directive supportive therapy for depression: A meta-analytic review. *Journal of Affective Disorders*, *349*, 452–461. https://doi.org/10.1016/j.jad.2024.01.073

Cuijpers, P., Miguel, C., Ciharova, M., Harrer, M., Basic, D., Cristea, I. A., ... & Karyotaki, E. (2024). Absolute and relative outcomes of psychotherapies for eight mental disorders: A systematic review and meta-analysis. *World Psychiatry*, *23*(2), 267–275. https://doi.org//10.1002/wps.21203

Cuijpers, P., Miguel, C., Harrer, M., Ciharova, M., & Karyotaki, E. (2024). The overestimation of the effect sizes of psychotherapies for depression in waitlist controlled trials: a meta-analytic comparison with usual care controlled trials. *Epidemiology and Psychiatric Sciences*, *33*, Article e56. https://doi.org/10.1017/S2045796024000611

Cuijpers, P., Miguel, C., Harrer, M., Plessen, C. Y., Ciharova, M., Ebert, D., & Karyotaki, E. (2025). *Database of depression psychotherapy trials with control conditions. Part of the Metapsy project* (Version 24.0.2) [Data set]. https://doi.org/10.5281/zenodo.7254845

†Cuijpers, P., Sijbrandij, M., Koole, S., Huibers, M., Berking, M., & Andersson, G. (2014). Psychological treatment of generalized anxiety disorder: A meta-analysis. *Clinical Psychology Review*, *34*(2), 130–140. https://doi.org/10.1016/j.cpr.2014.01.002

Cuijpers, P., Turner, E. H., Koole, S. L., van Dijke, A., & Smit, F. (2014). What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depression and Anxiety*, *31*(5), 374–378. https://doi.org/10.1002/da.22249

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: A meta-analysis of comparative outcome studies. *Journal of Consulting and Clinical Psychology*, *76*(6), 909–922. https://doi.org/10.1037/a0013075

Cuijpers, P., van Straten, A., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). The effects of psychotherapy for adult depression are overestimated: A meta-analysis of study quality and effect size. *Psychological Medicine*, *40*(2), 211–223. https://doi.org/10.1017/S0033291709006114

Cuijpers, P., Veen, S. C. V., Sijbrandij, M., Yoder, W., & Cristea, I. A. (2020). Eye movement desensitization and reprocessing for mental health problems: A systematic review and meta-analysis. *Cognitive Behaviour Therapy, 49*, 165–180.

†Currier, J. M., Holland, J. M., & Neimeyer, R. A. (2010). Do CBT-based interventions alleviate distress following bereavement? A review of the current evidence. *International Journal of Cognitive Therapy*, *3*(1), 77–93. https://doi.org/10.1521/ijct.2010.3.1.77

*Daniel, S. I. F., Poulsen, S., & Lunn, S. (2016). Client attachment in a randomized clinical trial of psychoanalytic and cognitive-behavioral psychotherapy for bulimia nervosa: outcome moderation and change. *Psychotherapy, 53*(2), 174–184. https://doi.org/10.1037/pst0000046

Davidson, P. R., & Parker, K. C. H. (2001). Eye Movement Desensitization and Reprocessing (EMDR): A meta-analysis. *Journal of Consulting and Clinical Psychology*, *69*(2), 305–316. https://doi.org/10.1037/0022-006X.69.2.305

de Lorgeril, M., Salen, P., Martin, J. L., Monjaud, I., Boucher, P., & Mamelle, N. (1998). Mediterranean dietary pattern in a randomized trial: Prolonged survival and possible reduced cancer rate. *Archives of Internal Medicine*, *158*(11), 1181–1187. https://doi.org/10.1001/archinte.158.11.1181

*de Ruiter, C., Rijken, H., Kraaimaat, F., & Garssen, B. (1989). Breathing retraining, exposure and a combination of both, in the treatment of panic disorder with agoraphobia. *Behaviour Research and Therapy*, *27*(6), 647–655. https://doi.org/10.1016/0005-7967(89)90148-4

*Deckersbach, T., Rauch, S., Buhlmann, U., & Wilhelm, S. (2006). Habit reversal versus supportive psychotherapy in Tourette's disorder: A randomized controlled trial and predictors of treatment response. *Behaviour Research and Therapy*, *44*(8), 1079–1090. https://doi.org/10.1016/j.brat.2005.08.007

Demicheli, V., Jefferson, T., Ferroni, E., Rivetti, A., & Di Pietrantonj, C. (2018). Vaccines for preventing influenza in healthy adults. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD001269.pub6

Derry, S., Wiffen, P. J., Moore, R. A., & Bendtsen, L. (2015). Ibuprofen for acute treatment of episodic tension-type headache in adults. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD011474.pub2

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

DerSimonian, R., & Laird, N. (2015). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, *45*(A), 139–145. https://doi.org/10.1016/j.cct.2015.09.002

*Dimidjian, S., Hollon, S. D., Dobson, K. S., Schmaling, K. B., Kohlenberg, R. J., Addis, M. E., Gallop, R., McGlinchey, J. B., Markley, D. K., Gollan, J. K., Atkins, D. C., Dunner, D. L., & Jacobson, N. S. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology*, *74*(4), 658–670. https://doi.org/10.1037/0022-006X.74.4.658

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *57*(3), 414–419. https://doi.org/10.1037/0022-006X.57.3.414

*Driessen, E., Van, H. L., Don, F. J., Peen, J., Kool, S., Westra, D., Hendriksen, M., Schoevers, R. A., Cuijpers, P., Twisk, J. W. R., & Dekker, J. J. M. (2013). The efficacy of cognitive-behavioral therapy and psychodynamic therapy in the outpatient treatment of major depression: A randomized clinical trial. *The American Journal of Psychiatry*, *170*(9), 1041–1050. https://doi.org/10.1176/appi.ajp.2013.12070899

Dubois, P. (1906). Rational psycho-therapeutics. *The British Medical Journal, 2*(2387), 767. http://www.dx.doi.org/10.1136/bmj.2.2387.741

*Dugas, M. J., Brillon, P., Savard, P., Turcotte, J., Gaudet, A., Ladouceur, R., Leblanc, R., & Gervais, N. J. (2010). A randomized clinical trial of cognitive-behavioral therapy and applied relaxation for adults with generalized anxiety disorder. *Behavior Therapy*, *41*(1), 46–58. https://doi.org/10.1016/j.beth.2008.12.004

*Durham, R. C., Murphy, T., Allan, T., Richard, K., Treliving, L. R., & Fenton, G. W. (1994). Cognitive therapy, analytic psychotherapy and anxiety management training for generalised anxiety disorder. *British Journal of Psychiatry*, *165*(3), 315–323. https://doi.org/10.1192/bjp.165.3.31

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

*Ehlers, A., Hackmann, A., Grey, N., Wild, J., Liness, S., Albert, I., Deale, A., Stott, R., & Clark, D. M. (2014). A randomized controlled trial of 7-day intensive and standard weekly cognitive therapy for ptsd and emotion-focused supportive therapy. *The American Journal of Psychiatry*, *171*(3), 294–304. https://doi.org/10.1176/appi.ajp.2013.13040552

†Ekers, D., Richards, D., & Gilbody, S. (2008). A meta-analysis of randomized trials of behavioural treatment of depression. *Psychological Medicine*, *38*(5), 611–623. https://doi.org/10.1017/S0033291707001614

*Ellison, J. A., Greenberg, L. S., Goldman, R. N., & Angus, L. (2009). Maintenance of gains following experiential therapies for depression. *Journal of Consulting and Clinical Psychology*, *77*(1), 103–112. https://doi.org/10.1037/a0014653

*Emmelkamp, P. M. G., & Beens, H. (1991). Cognitive therapy with obsessive-compulsive disorder: A comparative evaluation. *Behaviour Research and Therapy*, *29*(3), 293–300. https://doi.org/10.1016/0005-7967(91)90120-R

*Emmelkamp, P. M. G., Benner, A., Kuipers, A., Feiertag, G. A., Koster, H. C., & van Apeldoorn, F. J. (2006). Comparison of brief dynamic and cognitive–behavioural therapies in avoidant personality disorder. *British Journal of Psychiatry*, *189*(1), 60–64. https://doi.org/10.1192/bjp.bp.105.012153

*Emmelkamp, P. M. G., Krijn, M., Hulsbosch, A. M., de Vries, S., Schuemie, M. J., & van der Mast, C. A. P. G. (2002). Virtual reality treatment versus exposure in vivo: a comparative evaluation in acrophobia. *Behaviour Research and Therapy*, *40*(5), 509–516. https://doi.org/10.1016/S0005-7967(01)00023-7

*Emmelkamp, P. M. G., Visser, S., & Hoekstra, R. J. (1988). Cognitive therapy vs exposure in vivo in the treatment of obsessive-compulsives. *Cognitive Therapy and Research*, *12*(1), 103–114. https://doi.org/10.1007/BF01172784

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology, 16*(5), 319–324. https://www.dx.doi.org/10.1037/h0063633

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33*(5), 517. https://www.dx.doi.org/10.1037//0003-066x.33.5.517.a

Eysenck, H. J. (1994). Systematic reviews: Meta-analysis and its problems. *BMJ, 309*(6957), 789–792. https://doi.org/10.1136/bmj.309.6957.789

Eysenck, H. J., & Martin, E. (Eds.). (1987). *Theoretical foundations of behavior therapy*. Plenum.

*Fairburn, C. G., Bailey-Straebler, S., Basden, S., Doll, H. A., Jones, R., Murphy, R., O'Connor, M. E., & Cooper, Z. (2015). A transdiagnostic comparison of enhanced cognitive behaviour therapy (CBT-E) and interpersonal psychotherapy in the treatment of eating disorders. *Behaviour Research and Therapy*, *70*, 64–71. https://doi.org/10.1016/j.brat.2015.04.010

Falkenström, F., Granström, F., & Holmqvist, R. (2013). Therapeutic alliance predicts symptomatic improvement session by session. *Journal of Counseling Psychology*, *60*(3), 317–328. https://doi.org/10.1037/a0032258

*Farahimanesh, S., Moradi, A., Sadeghi, M., & Jobson, L. (2021). Comparing the efficacy of Competitive Memory Training (COMET) and MEmory Specificity Training (MEST) on posttraumatic stress disorder among newly diagnosed cancer patients. *Cognitive Therapy and Research*, *45*(5), 918–928. https://doi.org/10.1007/s10608-020-10175-4

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). John Wiley & Sons.

†Flückiger, C., Carratta, K., Del Re, A. C., Probst, G., Vîslă, A., Gómez Penedo, J. M., & Wampold, B. E. (2022). The relative efficacy of bona fide cognitive behavioral therapy and applied relaxation for generalized anxiety disorder at follow-up: A longitudinal multilevel meta-analysis. *Journal of Consulting and Clinical Psychology*, *90*(4), 339–352. https://doi.org/10.1037/ccp0000717

*Flückiger, C., Forrer, L., Schnider, B., Bättig, I., Bodenmann, G., & Zinbarg, R. E. (2016). A Single-blinded, randomized clinical trial of how to implement an evidence-based treatment for generalized anxiety disorder [IMPLEMENT] — effects of three different

strategies of implementation. *EBioMedicine*, *3*, 163–171. https://doi.org/10.1016/j.ebiom.2015.11.049

Foa, E. B., & Meadows, E. A. (1997). Psychosocial treatments for posttraumatic stress disorder: A critical review. *Annual Review of Psychology*, *48*, 449–480. https://doi.org/10.1146/annurev.psych.48.1.449

*Foa, E. B., Dancu, C. V., Hembree, E. A., Jaycox, L. H., Meadows, E. A., & Street, G. P. (1999). A comparison of exposure therapy, stress inoculation training, and their combination for reducing posttraumatic stress disorder in female assault victims. *Journal of Consulting and Clinical Psychology*, *67*(2), 194–200. https://doi.org/10.1037/0022-006X.67.2.194

*Foa, E. B., Hembree, E. A., Cahill, S. P., Rauch, S. A. M., Riggs, D. S., Feeny, N. C., & Yadin, E. (2005). Randomized trial of prolonged exposure for posttraumatic stress disorder with and without cognitive restructuring: Outcome at academic and community clinics. *Journal of Consulting and Clinical Psychology*, *73*(5), 953–964. https://doi.org/10.1037/0022-006X.73.5.953

Foa, E. B., Keane, T. M., & Friedman, M. J. (2000). Guidelines for treatment of PTSD. *Journal of Traumatic Stress*, *13*(4), 539–588. https://doi.org/10.1023/A:1007802031411

*Foa, E. B., Rothbaum, B. O., Riggs, D. S., & Murdock, T. B. (1991). Treatment of posttraumatic stress disorder in rape victims: A comparison between cognitive-behavioral procedures and counseling. *Journal of Consulting and Clinical Psychology*, *59*(5), 715–723. https://doi.org/10.1037/0022-006X.59.5.715

Foa, E. B., Zoellner, L. A., Feeny, N. C., Hembree, E. A., & Alvarez-Conrad, J. (2002). Does imaginal exposure exacerbate PTSD symptoms? *Journal of Consulting and Clinical Psychology*, *70*(4), 1022–1028. https://doi.org/10.1037/0022-006X.70.4.1022

*Fonagy, P., Lemma, A., Target, M., O'Keeffe, S., Constantinou, M. P., Ventura Wurman, T., Luyten, P., Allison, E., Roth, A., Cape, J., & Pilling, S. (2020). Dynamic interpersonal therapy for moderate to severe depression: A pilot randomized controlled and feasibility trial. *Psychological Medicine*, *50*(6), 1010–1019. https://doi.org/10.1017/S0033291719000928

Frank, J. D. (1961). *Persuasion and healing: A comparative study of psychotherapy* (1st ed.). The Johns Hopkins Press.

Frank, J. D. (1971). Therapeutic factors in psychotherapy. *American Journal of Psychotherapy*, *25*(3), 350-361.

*Freeman, C. P. L., Barry, F., Dunkeld-Turnbull, J., & Henderson, A. (1988). Controlled trial of psychotherapy for bulimia nervosa. *BMJ*, *296*(6621), 521–525. https://doi.org/10.1136/bmj.296.6621.521

Furukawa, T. A., & Leucht, S. (2011). How to obtain NNT from Cohen's *d*: Comparison of two methods. *PLoS ONE, 6*(4), 1–5. https://doi.org/10.1371/ journal.pone.0019070

†Furukawa, T. A., Shinohara, K., Sahker, E., Karyotaki, E., Miguel, C., Ciharova, M., Bockting, C. L. H., Breedvelt, J. J. F., Tajika, A., Imai, H., Ostinelli, E. G., Sakata, M., Toyomoto, R., Kishimoto, S., Ito, M., Furukawa, Y., Cipriani, A., Hollon, S. D., & Cuijpers, P. (2021). Initial treatment choices to achieve sustained response in major depression: a systematic review and network meta-analysis. *World Psychiatry, 20*(3), 387–396. https://doi.org/10.1002/wps.20906

Germer, S., Weyrich, V., Bräscher, A.-K., Mütze, K., & Witthöft, M. (2022). Does Practice Really Make Perfect? A Longitudinal Analysis of the Relationship Between Therapist Experience and Therapy Outcome: A Replication of Goldberg, Rousmaniere, et al. (2016). *Journal of Counseling Psychology*, *69*(5), 745–754. https://doi.org/10.1037/cou0000608

*Ghaderi, A. (2006). Does individualization matter? A randomized trial of standardized (focused) versus individualized (broad) cognitive behavior therapy for bulimia nervosa. *Behaviour Research and Therapy*, *44*(2), 273–288. https://doi.org/10.1016/j.brat.2005.02.004

*Giesen-Bloo, J., van Dyck, R., Spinhoven, P., van Tilburg, W., Dirksen, C., van Asselt, T., Kremers, I., Nadort, M., & Arntz, A. (2006). Outpatient psychotherapy for borderline personality disorder: Randomized trial of schema-focused therapy vs transference-focused psychotherapy. *Archives of General Psychiatry*, *63*(6), 649–658. https://doi.org/10.1001/archpsyc.63.6.649

†Ginley, M. K., Pfund, R. A., Rash, C. J., & Zajac, K. (2021). Long-term efficacy of contingency management treatment based on objective indicators of abstinence from illicit substance use up to 1 year following treatment: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *89*(1), 58–71. https://doi.org/10.1037/ccp0000552

Glass, G. V (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3–8. https://doi.org/10.2307/1174772

Glass, G. V. (2000). *Meta-analysis at 25*. https://web.archive.org/web/20220303150718/http://www.gvglass.info/papers/meta25.html

*Gloster, A. T., Hauke, C., Höfler, M., Einsle, F., Fydrich, T., Hamm, A., Sthröhle, A., & Wittchen, H.-U. (2013). Long-term stability of cognitive behavioral therapy effects for panic disorder with agoraphobia: A two-year follow-up study. *Behaviour Research and Therapy*, *51*(12), 830–839. https://doi.org/10.1016/j.brat.2013.09.009

Goldberg, S. B., Tucker, R. P., Greene, P. A., Davidson, R. J., Wampold, B. E., Kearney, D. J., & Simpson, T. L. (2018). Mindfulness-based interventions for psychiatric disorders: A systematic review and meta-analysis. *Clinical Psychology Review*, *59*, 52–60. https://doi.org/10.1016/j.cpr.2017.10.011

Goldfried, M. R. (1980). Toward the delineation of therapeutic change principles. *The American Psychologist*, *35*(11), 991–999. https://doi.org/10.1037/0003-066X.35.11.991

Goldfried, M. R., & Kent, R. N. (1972). Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin*, *77*(6), 409-420. https://doi.org/10.1037/h0032714

Goldfried, M. R., & Wolfe, B. E. (1996). Psychotherapy practice and research: Repairing a strained relationship. *American Psychologist*, *51*(10), 1007–1016. https://doi.org/10.1037/0003-066X.51.10.1007

Goldstein, L. A., Mandel, A. D. A., DeRubeis, R. J., & Strunk, D. R. (2020). Outcomes, skill acquisition, and the alliance: Similarities and differences between clinical trial and student therapists. *Behaviour Research and Therapy, 129*, Article 103608. https://doi.org/10.1016/j.brat.2020.103608

Goleman, D. (1990, March 6). As a therapist, Freud fell short, scholars find. *The New York Times*, pp. C1, C12.

Gosselin, R. A., Roberts, I., & Gillespie, W. J. (2004). Antibiotics for preventing infection in open limb fractures. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD003764.pub2

Graves, T. A., Tabri, N., Thompson-Brenner, H., Franko, D. L., Eddy, K. T., Bourion-Bedes, S., Brown, A., Constantino, M. J., Flückiger, C., Forsberg, S., Isserlin, L., Couturier, J., Paulson Karlsson, G., Mander, J., Teufel, M., Mitchell, J. E., Crosby, R. D., Prestano, C., Satir, D. A., … Thomas, J. J. (2017). A meta-analysis of the relation between therapeutic alliance and treatment outcome in eating disorders. *The International Journal of Eating Disorders*, *50*(4), 323–340. https://doi.org/10.1002/eat.22672

*Greenberg, L., & Watson, J. (1998). Experiential therapy of depression: Differential effects of client-centered relationship conditions and process experiential interventions. *Psychotherapy Research*, *8*(2), 210–224. https://doi.org/10.1080/10503309812331332317

†Grenon, R., Carlucci, S., Brugnera, A., Schwartze, D., Hammond, N., Ivanova, I., Mcquaid, N., Proulx, G., & Tasca, G. A. (2019). Psychotherapy for eating disorders: A meta-analysis of direct comparisons. *Psychotherapy Research*, *29*(7), 833–845. https://doi.org/10.1080/10503307.2018.1489162

*grosse Holtforth, M., Krieger, T., Zimmermann, J., Altenstein-Yamanaka, D., Dörig, N., Meisch, L., & Hayes, A. M. (2019). A randomized-controlled trial of cognitive-behavioral therapy for depression with integrated techniques from emotion-focused and exposure therapies. *Psychotherapy Research*, *29*(1), 30–44. https://doi.org/10.1080/10503307.2017.1397796

†Gupta, A. E. (2021). *The efficacy of EMDR versus TF-CBT for the treatment of PTSD* (Publication No. 28773877) [Doctoral dissertation, California Southern University]. ProQuest Dissertations & Theses.

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, *2*(3), 109–112. https://doi.org/10.4103/2229-3485.83221

†Halicka, M., Parkhouse, T. L., Webster, K., Spiga, F., Hines, L. A., Freeman, T. P., Sanghera, S., Dawson, S., Paterson, C., Savović, J., Higgins, J. P. T., & Caldwell, D. M. (2025). Effectiveness and safety of psychosocial interventions for the treatment of cannabis use disorder: A systematic review and meta-analysis. *Addiction*. https://doi.org/10.1111/add.70084

*Hardy, G. E., Barkham, M., Shapiro, D. A., Stiles, W. B., Rees, A., & Reynolds, S. (1995). Impact of Cluster C personality disorders on outcomes of contrasting brief psychotherapies for depression. *Journal of Consulting and Clinical Psychology.*, *63*(6), 997–1004. https://doi.org/10.1037/0022-006X.63.6.997

Harrer, M., Miguel, C., van Ballegooijen, W., Ciharova, M., Plessen, C. Y., Kuper, P., Sprenger, A. A., Buntrock, C., Papola, D., Cristea, I. A., de Ponti, N., Bašić, Đ., Pauley, D., Driessen, E., Quero, S., Grimaldos, J., Buendía, S. F., Botella, C., Hamblen, J. L., & Schnurr, P. P. (2025). Effectiveness of psychotherapy: Synthesis of a "meta-analytic research domain" across world regions and 12 mental health problems. *Psychological Bulletin.*, *151*(5), 600–667. https://doi.org/10.1037/bul0000465

Hayes, S. C., Follette, V. M., & Linehan, M. (2004). *Mindfulness and acceptance: Expanding the cognitive-behavioral tradition*. Guilford Press.

Hayes, S. C., & Hofmann, S. G. (2017). The third wave of cognitive behavioral therapy and the rise of process-based care. *World Psychiatry*, *16*(3), 245–246. https://doi.org/10.1002/wps.20442

Hayes, S. C., & Hofmann, S. G. (Eds.). (2018). *Process-based CBT: The science and core clinical competencies of cognitive behavioral therapy*. New Harbinger Publications.

*Hayes-Skelton, S. A., Roemer, L., & Orsillo, S. M. (2013). A randomized clinical trial comparing an acceptance-based behavior therapy to applied relaxation for generalized anxiety disorder. *Journal of Consulting and Clinical Psychology.*, *81*(5), 761–773. https://doi.org/10.1037/a0032871

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators . *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217. https://doi.org/10.1037/1082-989X.6.3.203

Hedges, L. V., & Vevea, J. L. (1998). Fixed and random-effects models in meta-analysis. Psychological Methods, 3(4), 486–504.

*Hellström, K., Fellenius, J., & öst, L.-G. (1996). One versus five sessions of applied tension in the treatment of blood phobia. *Behaviour Research and Therapy.*, *34*(2), 101–112. https://doi.org/10.1016/0005-7967(95)00060-7

*Hemanny, C., Carvalho, C., Maia, N., Reis, D., Botelho, A. C., Bonavides, D., Seixas, C., & de Oliveira, I. R. (2020). Efficacy of trial-based cognitive therapy, behavioral activation and treatment as usual in the treatment of major depressive disorder: Preliminary findings

from a randomized clinical trial. *CNS Spectrums*, *25*(4), 535–544. https://doi.org/10.1017/S1092852919001457

*Hersen, M., Himmelhoch, J. M., Thase, M. E., & Bellack, A. S. (1984). Effects of social skill training, amitriptyline, and psychotherapy in unipolar depressed women. *Behavior Therapy.*, *15*(1), 21–40. https://doi.org/10.1016/S0005-7894(84)80039-8

*Hien, D. A., Cohen, L. R., Miele, G. M., Litt, L. C., & Capstick, C. (2004). Promising treatments for women with comorbid ptsd and substance use disorders. *The American Journal of Psychiatry.*, *161*(8), 1426–1432. https://doi.org/10.1176/appi.ajp.161.8.1426

*Hien, D. A., Smith, K. Z., Owens, M., López-Castro, T., Ruglass, L. M., & Papini, S. (2018). Lagged effects of substance use on PTSD severity in a randomized controlled trial with modified prolonged exposure and relapse prevention. *Journal of Consulting and Clinical Psychology.*, *86*(10), 810–819. https://doi.org/10.1037/ccp0000345

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology*, *37*(5), 1158–1160. https://doi.org/10.1093/ije/dyn204

Higgins, J. P. T., & Green, S. (Eds.) (2011). *Cochrane handbook for systematic reviews of interventions* (Version. 5.1). The Cochrane Collaboration. http://www.handbook.cochrane.org.

Higgins, J. P. T., & 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. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119536604

Higgins, J. P. T., 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.327.7414.557Higgins et al., 2003

†Ho, M., & Lee, C. (2012). Cognitive behaviour therapy versus eye movement desensitization and reprocessing for post-traumatic disorder – is it all in the homework then? *European Review of Applied Psychology*, *62*(4), 253–260. https://doi.org/10.1016/j.erap.2012.08.001

*Hoffart, A., Øktedalen, T., Langkaas, T. F., & Wampold, B. E. (2013). Alliance and outcome in varying imagery procedures for PTSD: A study of within-person processes. *Journal of Counseling Psychology.*, *60*(4), 471–482. https://doi.org/10.1037/a0033604

Hofmann, S. G., & Hayes, S. C. (2019). The future of intervention science: Process-based therapy. *Clinical Psychological Science*, *7*(1), 37–50. https://doi.org/10.1177/2167702618772296

Hofmann, S. G., Kasch, C., & Reis, A. (2025). Effect sizes of randomized-controlled studies of cognitive behavioral therapy for anxiety disorders over the past 30 years. *Clinical Psychology Review*, *117*, Article 102553. https://doi.org/10.1016/j.cpr.2025.102553

Hopewell, S., Clarke, M., and 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*. John Wiley & Sons.

*Hopko, D. R., Armento, M. E. A., Robertson, S. M. C., Ryba, M. M., Carvalho, J. P., Colman, L. K., Mullane, C., Gawrysiak, M., Bell, J. L., McNulty, J. K., & Lejuez, C. W. (2011). Brief behavioral activation and problem-solving therapy for depressed breast cancer

patients: Randomized trial. *Journal of Consulting and Clinical Psychology.*, *79*(6), 834–849. https://doi.org/10.1037/a0025450

†Hoppen, T. H., Lindemann, A. S., & Morina, N. (2022). Safety of psychological interventions for adult post-traumatic stress disorder: Meta-analysis on the incidence and relative risk of deterioration, adverse events and serious adverse events. *The British Journal of Psychiatry.*, *221*(5), 658–667. https://doi.org/10.1192/bjp.2022.111

*Horst, F., Den Oudsten, B., Zijlstra, W., de Jongh, A., Lobbestael, J., & De Vries, J. (2017). Cognitive behavioral therapy vs. eye movement desensitization and reprocessing for treating panic disorder: A randomized controlled trial. *Frontiers in Psychology.*, *8*. https://doi.org/10.3389/fpsyg.2017.01409

Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy, 48*(1), 9–16. https://doi.org/10.1037/a0022186

*Hoyer, J., Beesdo, K., Gloster, A. T., Runge, J., Höfler, M., & Becker, E. S. (2009). Worry exposure versus applied relaxation in the treatment of generalized anxiety disorder. *Psychotherapy and Psychosomatics.*, *78*(2), 106–115. https://doi.org/10.1159/000201936

*Huber, D., Henrich, G., Gastner, J., & Klug, G. (2012). Must all have prizes? The Munich Psychotherapy Study. In R. A. Levy, H. Kächele, & J. S. Ablon (Eds.), *Psychodynamic Psychotherapy Research* (pp. 51–69). Humana Press. https://doi.org/10.1007/978-1-60761-792-1_4

Hunot, V., Churchill, R., Teixeira, V., & Silva de Lima, M. (2007). Psychological therapies for generalised anxiety disorder. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD001848.pub4

Hunsley, J., & Di Giulio, G. (2002). Dodo bird, phoenix, or urban legend? The question of psychotherapy equivalence. *The Scientific Review of Mental Health Practice, 1*(1), 11–22.

Hunsley, J., & Westmacott, R. (2007). Interpreting the magnitude of the placebo effect: Mountain or molehill? *Journal of Clinical Psychology*, *63*(4), 391–399. https://doi.org/10.1002/jclp.20352

Hunsley, J., Elliott, K., & Therrien, Z. (2014). The Efficacy and Effectiveness of Psychological Treatments for Mood, Anxiety, and Related Disorders. *Canadian Psychology*, *55*(3), 161–176. https://doi.org/10.1037/a0036933IntHout et al., 2016

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*(7), Article e010247. https://doi.org/10.1136/bmjopen-2015-010247

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

*Ito, L. M., Araujo, L. A. D., Tess, V. L. C., Barros-Neto, T. P. D., Asbahr, F. R., & Marks, I. (2001). Self-exposure therapy for panic disorder with agoraphobia: Randomised controlled study of external v. interoceptive self-exposure. *British Journal of Psychiatry*, *178*(4), 331–336. https://doi.org/10.1192/bjp.178.4.331

*Jackson, J. B., Pietrabissa, G., Rossi, A., Manzoni, G. M., & Castelnuovo, G. (2018). Brief strategic therapy and cognitive behavioral therapy for women with binge eating disorder and comorbid obesity: A randomized clinical trial one-year follow-up. *Journal of Consulting and Clinical Psychology*, *86*(8), 688–701. https://doi.org/10.1037/ccp0000313

*Jacobson, N. S., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., Gortner, E., & Prince, S. E. (1996). A component analysis of cognitive-behavioral treatment for depression. *Journal of Consulting and Clinical Psychology*, *64*(2), 295–304. https://doi.org/10.1037/0022-006X.64.2.295

James, W. (2013). *Pragmatism: A new name for some old ways of thinking*. Project Gutenberg. https://www.gutenberg.org/files/5116/5116-h/5116-h.htm#link2H_4_0004 (Original work published 1907)

†Jeong, H., Yim, H. W., Lee, S.-Y., Potenza, M. N., & Kim, N.-J. (2023). Effectiveness of psychotherapy on prevention of suicidal re-attempts in psychiatric emergencies: A systematic review and network meta-analysis of randomized controlled trials. *Psychotherapy and Psychosomatics.*, *92*(3), 152–161. https://doi.org/10.1159/000529753

Jiménez-Orenga, N., Miguel, C., González-Robles, A., Fernández-Álvarez, J., Grimaldos, J., Bretón-López, J., Botella, C., Cuijpers, P., García-Palacios, A., Papola, D., Quero, S., Riper, H., & Díaz-García, A. (2025). Transdiagnostic psychological interventions for emotional disorders: A comprehensive meta-analysis. *Journal of Affective Disorders.*, *388*. https://doi.org/10.1016/j.jad.2025.119537

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin*, *141*(4), 747–768. https://doi.org/10.1037/bul0000015

†Jones, C., Hacker, D., Meaden, A., Cormac, I., Irving, C. B., Xia, J., Zhao, S., Shi, C., & Chen, J. (2018). Cognitive behavioural therapy plus standard care versus standard care plus other psychosocial treatments for people with schizophrenia. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD008712.pub3

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature (London)*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

*Kampmann, I. L., Emmelkamp, P. M., Hartanto, D., Brinkman, W.-P., Zijlstra, B. J., & Morina, N. (2016). Exposure to virtual social interactions in the treatment of social anxiety disorder: A randomized controlled trial. *Behaviour Research and Therapy.*, *77*, 147–156. https://doi.org/10.1016/j.brat.2015.12.016

Kataoka, Y., Takayama, T., Yoshimura, K., So, R., Tsujimoto, Y., Yamagishi, Y., Takagi, S., Furukawa, Y., Sakata, M., Bašić, Đ., Cipriani, A., Cuijpers, P., Karyotaki, E., Harrer, M., Leucht, S., Homiar, A., Ostinelli, E. G., Miguel, C., Rodolico, A., & Furukawa, T. A. (2025). Automating the data extraction process for systematic reviews using GPT-4o and o3. *Research Synthesis Methods*, 1–21. https://doi.org/10.1017/rsm.2025.10030

*Keijsers, G. P. J., Maas, J., van Opdorp, A., & van Minnen, A. (2016). Addressing self-control cognitions in the treatment of trichotillomania: A randomized controlled trial comparing cognitive therapy to behaviour therapy. *Cognitive Therapy and Research.*, *40*(4), 522–531. https://doi.org/10.1007/s10608-016-9754-4

*Kiosses, D. N., Arean, P. A., Teri, L., & Alexopoulos, G. S. (2010). Home-delivered problem adaptation therapy (path) for depressed, cognitively impaired, disabled elders: A preliminary study. *American Journal of Geriatric Psychiatry.*, *18*(11), 988–998. https://doi.org/10.1097/JGP.0b013e3181d6947d

†Kivlighan, D. M., Goldberg, S. B., Abbas, M., Pace, B. T., Yulish, N. E., Thomas, J. G., Cullen, M. M., Flückiger, C., & Wampold, B. E. (2015). The enduring effects of psychodynamic treatments vis-à-vis alternative treatments: A multilevel longitudinal meta-analysis. *Clinical Psychology Review.*, *40*, 1–14. https://doi.org/10.1016/j.cpr.2015.05.003

Klatte, R., Strauss, B., Flückiger, C., Färber, F., & Rosendahl, J. (2023). Defining and assessing adverse events and harmful effects in psychotherapy study protocols: A systematic review. *Psychotherapy*, *60*(1), 130–148. https://doi.org/10.1037/pst0000359

*Klein, D. N., Leon, A. C., Li, C., D'Zurilla, T. J., Black, S. R., Vivian, D., Dowling, F., Arnow, B. A., Manber, R., Markowitz, J. C., & Kocsis, J. H. (2011). Social problem solving and depressive symptoms over time: A randomized clinical trial of cognitive-behavioral analysis system of psychotherapy, brief supportive psychotherapy, and pharmacotherapy. *Journal of Consulting and Clinical Psychology.*, *79*(3), 342–352. https://doi.org/10.1037/a0023208

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

*Koszycki, D., Bisserbe, J.-C., Blier, P., Bradwejn, J., & Markowitz, J. (2012). Interpersonal psychotherapy versus brief supportive therapy for depressed infertile women: first pilot randomized controlled trial. *Archives of Women's Mental Health.*, *15*(3), 193–201. https://doi.org/10.1007/s00737-012-0277-z

*Koszycki, D., Raab, K., Aldosary, F., & Bradwejn, J. (2010). A multifaith spiritually based intervention for generalized anxiety disorder: a pilot randomized trial. *Journal of Clinical Psychology.*, *66*(4), 430–441. https://doi.org/10.1002/jclp.20663

†Kotova, E. (2005). *A meta -analysis of interpersonal psychotherapy* (Publication No. 3177607) [Doctoral dissertation, University of Toledo]. ProQuest Dissertations & Theses.

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological psychiatry*, *59*(11), 990–996. https://doi.org/10.1016/j.biopsych.2005.09.014

*Kramer, U., Kolly, S., Berthoud, L., Keller, S., Preisig, M., Caspar, F., Berger, T., de Roten, Y., Marquet, P., & Despland, J.-N. (2014). Effects of motive-oriented therapeutic relationship in a ten-session general psychiatric treatment of borderline personality disorder: a randomized controlled trial. *Psychotherapy and Psychosomatics.*, *83*(3), 176–186. https://doi.org/10.1159/000358528

*Kunze, A. E., Arntz, A., Morina, N., Kindt, M., & Lancee, J. (2017). Efficacy of imagery rescripting and imaginal exposure for nightmares: A randomized wait-list controlled trial. *Behaviour Research and Therapy.*, *97*, 14–25. https://doi.org/10.1016/j.brat.2017.06.005

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers* (Vol. 1). Cambridge University Press.

Lambert, M. J. (1976). Spontaneous remission in adult neurotic disorders: A revision and summary. *Psychological Bulletin*, *83*(1), 107–119. https://doi.org/10.1037/0033-2909.83.1.107

Lambert, M. J. (Ed.). (2013). *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed.). John Wiley & Sons.

Lambert, M. J., & Bergin, A. E. (1994). The effectiveness of psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 143–189). John Wiley & Sons.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Vermeersch, D. A., Clouse, G. C., & Yanchar, S. C. (1996). The reliability and validity of the Outcome Questionnaire. *Clinical Psychology and Psychotherapy*, *3*(4), 249–258. https://doi.org/10.1002/(SICI)1099-0879(199612)3:4<249::AID-CPP106>3.0.CO;2-S

*Langkaas, T. F., Hoffart, A., Øktedalen, T., Ulvenes, P. G., Hembree, E. A., & Smucker, M. (2017). Exposure and non-fear emotions: A randomized controlled study of exposure-based and rescripting-based imagery in PTSD treatment. *Behaviour Research and Therapy.*, *97*, 33–42. https://doi.org/10.1016/j.brat.2017.06.007

†Lee, E. B., An, W., Levin, M. E., & Twohig, M. P. (2015). An initial meta-analysis of Acceptance and Commitment Therapy for treating substance use disorders. *Drug and Alcohol Dependence.*, *155*, 1–7. https://doi.org/10.1016/j.drugalcdep.2015.08.004

Leichsenring, F., Abbass, A., Hilsenroth, M. J., Leweke, F., Luyten, P., Keefe, J. R., … & Steinert, C. (2017). Biases in research: Risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychological Medicine*, *47*(6), 1000–1011. https://doi.org/10.1017/S003329171600324X

Leichsenring, F., Rabung, S., & Leibing, E. (2004). The efficacy of short-term psychodynamic psychotherapy in specific psychiatric disorders: A meta-analysis. *Archives of General Psychiatry, 61*(12), 1208–1216. https://doi.org/10.1001/archpsyc.61.12.1208

*Leichsenring, F., Salzer, S., Beutel, M. E., Herpertz, S., Hiller, W., Hoyer, J., Huesing, J., Joraschky, P., Nolting, B., Poehlmann, K., Ritter, V., Stangier, U., Strauss, B.,

Stuhldreher, N., Tefikow, S., Teismann, T., Willutzki, U., Wiltink, J., & Leibing, E. (2013). Psychodynamic therapy and cognitive-behavioral therapy in social anxiety disorder: A multicenter randomized controlled trial. *The American Journal of Psychiatry.*, *170*(7), 759–767. https://doi.org/10.1176/appi.ajp.2013.12081125

*Leichsenring, F., Salzer, S., Jaeger, U., Kächele, H., Kreische, R., Leweke, F., Rüger, U., Winkelbach, C., & Leibing, E. (2009). Short-term psychodynamic psychotherapy and cognitive-behavioral therapy in generalized anxiety disorder: A randomized, controlled trial. *The American Journal of Psychiatry.*, *166*(8), 875–881. https://doi.org/10.1176/appi.ajp.2009.09030441

*Lemmens, L. H. J. M., Arntz, A., Peeters, F., Hollon, S. D., Roefs, A., & Huibers, M. J. H. (2015). Clinical effectiveness of cognitive therapy v. interpersonal psychotherapy for depression: results of a randomized controlled trial. *Psychological Medicine.*, *45*(10), 2095–2110. https://doi.org/10.1017/S0033291715000033

Leucht, S., Helfer, B., Gartlehner, G., & Davis, J. M. (2015). How effective are common medications: a perspective based on meta-analyses of major drugs. *BMC Medicine*, *13*(1), Article 253. https://doi.org/10.1186/s12916-015-0494-1

†Lewis, C., Roberts, N. P., Andrew, M., Starling, E., & Bisson, J. I. (2020). Psychological therapies for post-traumatic stress disorder in adults: systematic review and meta-analysis. *European Journal of Psychotraumatology.*, *11*(1), Article 1729633. https://doi.org/10.1080/20008198.2020.1729633

Leykin, Y., & DeRubeis, R. J. (2009). Allegiance in psychotherapy outcome research: Separating association from bias. *Clinical Psychology: Science and Practice*, *16*(1), 54–65. https://doi.org/10.1111/j.1468-2850.2009.01143.x

†Linardon, J. (2018). Meta-analysis of the effects of cognitive-behavioral therapy on the core eating disorder maintaining mechanisms: implications for mechanisms of therapeutic change. *Cognitive Behaviour Therapy.*, *47*(2), 107–125. https://doi.org/10.1080/16506073.2018.1427785

†Linardon, J., & Brennan, L. (2017). The effects of cognitive-behavioral therapy for eating disorders on quality of life: A meta-analysis. *The International Journal of Eating Disorders.*, *50*(7), 715–730. https://doi.org/10.1002/eat.22719

†Linardon, J., Fitzsimmons-Craft, E. E., Brennan, L., Barillaro, M., & Wilfley, D. E. (2019). Dropout from interpersonal psychotherapy for mental health disorders: A systematic review and meta-analysis. *Psychotherapy Research*, *29*(7), 870–881. https://doi.org/10.1080/10503307.2018.1497215

Linehan, M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. Guilford Press.

*Lipsitz, J. D., Gur, M., Vermes, D., Petkova, E., Cheng, J., Miller, N., Laino, J., Liebowitz, M. R., & Fyer, A. J. (2008). A randomized trial of interpersonal therapy versus supportive therapy for social anxiety disorder. *Depression and Anxiety.*, *25*(6), 542–553. https://doi.org/10.1002/da.20364

Lohr, J. M., Gist, R., Deacon, B., Devilly, G. J., & Varker, T. (2015). Science-and non-science-based treatments for trauma-related stress disorders. In S. O. Lilienfeld, S. J. Lynn, J. M., Lohr, and C. Tavris (Eds.), *Science and Pseudoscience in Clinical Psychology* (pp. 277–321). Guilford Press.

Luborsky, L., Diguer, L., Seligman, D. A., Rosenthal, R., Krause, E. D., Johnson, S., Halperin, G., Bishop, M., Berman, J. S., & Schweizer, E. (1999). The researcher's own therapy

allegiances: A “wild card” in comparisons of treatment efficacy. *Clinical Psychology 6*(1), 95–106. https://doi.org/10.1093/clipsy.6.1.95

Luborsky, L., Singer, B., & Luborsky, L. (1975). Comparative studies of psychotherapies: Is it true that “everyone has won and all must have prizes”? *Archives of General Psychiatry*, *32*(8), 995–1008. https://doi.org/10.1001/archpsyc.1975.01760260059004

Magnusson, K. (2023). *A causal inference perspective on therapist effects*. PsyArXiv. https://doi.org/10.31234/osf.io/f7mvz

Mann, J. (1973). *Time-limited psychotherapy*. Harvard University Press.

†Marcus, D. K., O’Connell, D., Norris, A. L., & Sawaqdeh, A. (2014). Is the Dodo bird endangered in the 21st century? A meta-analysis of treatment comparison studies. *Clinical Psychology Review*, *34*(7), 519–530. https://doi.org/10.1016/j.cpr.2014.08.001

*Markowitz, J. C., Kocsis, J. H., Christos, P., Bleiberg, K., & Carlin, A. (2008). Pilot study of interpersonal psychotherapy versus supportive psychotherapy for dysthymic patients with secondary alcohol abuse or dependence. *The Journal of Nervous and Mental Disease*, *196*(6), 468–474. https://doi.org/10.1097/NMD.0b013e31817738f1

*Markowitz, J. C., Petkova, E., Neria, Y., Van Meter, P. E., Zhao, Y., Hembree, E., Lovell, K., Biyanova, T., & Marshall, R. D. (2015). Is exposure necessary? A randomized clinical trial of interpersonal psychotherapy for PTSD. *The American Journal of Psychiatry.*, *172*(5), 430–440. https://doi.org/10.1176/appi.ajp.2014.14070908

*Marks, I., Lovell, K., Noshirvani, H., Livanou, M., & Thrasher, S. (1998). Treatment of posttraumatic stress disorder by exposure and/or cognitive restructuring: A controlled

study. *Archives of General Psychiatry*, *55*(4), 317–325. https://doi.org/10.1001/archpsyc.55.4.317

Mason, L., Grey, N., & Veale, D. (2016). My therapist is a student? The impact of therapist experience and client severity on cognitive behavioural therapy outcomes for people with anxiety disorders. *Behavioural and Cognitive Psychotherapy*, *44*(2), 193–202. https://doi.org/10.1017/S1352465815000065

†Mavranezouli, I., Megnin-Viggars, O., Daly, C., Dias, S., Welton, N. J., Stockton, S., Bhutani, G., Grey, N., Leach, J., Greenberg, N., Katona, C., El-Leithy, S., & Pilling, S. (2020). Psychological treatments for post-traumatic stress disorder in adults: A network meta-analysis. *Psychological Medicine.*, *50*(4), 542–555. https://doi.org/10.1017/S0033291720000070

McCarthy, K. S., & Barber, J. P. (2009). The Multitheoretical List of Therapeutic Interventions (MULTI): Initial report. *Psychotherapy Research*, *19*(1), 96–113. https://doi.org/10.1080/10503300802524343

McCullough, L., Kuhn, N., Andrews, S., Kaplan, A., Wolf, J., & Lanza Hurley, C. (2003). *Treating affect phobia: A manual for short-term dynamic psychotherapy*. Guilford Press.

*McDonagh, A., Friedman, M., McHugo, G., Ford, J., Sengupta, A., Mueser, K., Demment, C. C., Fournier, D., Schnurr, P. P., & Descamps, M. (2005). Randomized trial of cognitive-behavioral therapy for chronic posttraumatic stress disorder in adult female survivors of childhood sexual abuse. *Journal of Consulting and Clinical Psychology.*, *73*(3), 515–524. https://doi.org/10.1037/0022-006X.73.3.515

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin.*, *111*(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361

*McIndoo, C., File, A., Preddy, T., Clark, C., & Hopko, D. (2016). Mindfulness-based therapy and behavioral activation: A randomized controlled trial with depressed college students. *Behaviour Research and Therapy.*, *77*, 118–128. https://doi.org/10.1016/j.brat.2015.12.012

*McKay, J. R., Lynch, K. G., Coviello, D., Morrison, R., Cary, M. S., Skalina, L., & Plebani, J. (2010). Randomized trial of continuing care enhancements for cocaine-dependent patients following initial engagement. *Journal of Consulting and Clinical Psychology.*, *78*(1), 111–120. https://doi.org/10.1037/a0018139

McQuaid, A., Sanatinia, R., Farquharson, L. et al. (2021). Patient experience of lasting negative effects of psychological interventions for anxiety and depression in secondary mental health care services: A national cross-sectional study. *BMC Psychiatry, 21*, Article 578. https://doi.org/10.1186/s12888-021-03588-2

*Mersch, P. P. A. (1995). The treatment of social phobia: The differential effectiveness of exposure in vivo and an integration of exposure in vivo, rational emotive therapy and social skills training. *Behaviour Research and Therapy.*, *33*(3), 259–269. https://doi.org/10.1016/0005-7967(94)00038-L

Messer, S. B., & Wampold, B. E. (2002). Let's face facts: Common factors are more potent than specific therapy ingredients. *Clinical Psychology: Science and Practice*, *9*(1), 21–25. https://doi.org/10.1093/clipsy.9.1.21

Miller, R. C., & Berman, J. S. (1983). The efficacy of cognitive behavior therapies: A quantitative review of the research evidence. *Psychological Bulletin, 94*(1), 39–53. https://doi.org/10.1037/0033-2909.94.1.39

Miller, S., Wampold, B., & Varhely, K. (2008). Direct comparisons of treatment modalities for youth disorders: a meta-analysis. *Psychotherapy Research*, *18*(1), 5–14. https://doi.org/10.1080/10503300701472131

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine, 6*(7), 1–6. https://doi.org/10.1371/ journal.pmed.1000097

*Morgenstern, J., Irwin, T. W., Wainberg, M. L., Parsons, J. T., Muench, F., Bux, D. A., Kahler, C. W., Marcus, S., & Schulz-Heik, J. (2007). A randomized controlled trial of goal choice interventions for alcohol use disorders among men who have sex with men. *Journal of Consulting and Clinical Psychology.*, *75*(1), 72–84. https://doi.org/10.1037/0022-006X.75.1.72

Morrissey, M. B. (2016). Meta-analysis of magnitudes, differences and variation in evolutionary parameters. *Journal of Evolutionary Biology*, *29*(10), 1882–1904. https://doi.org/10.1111/jeb.12950

Munder, T., Flückiger, C., Gerger, H., Wampold, B. E., & Barth, J. (2012). Is the allegiance effect an epiphenomenon of true efficacy differences between treatments? A meta-analysis. *Journal of Counseling Psychology*, *59*(4), 631–637. https://doi.org/10.1037/a0029571

Munder, T., Gerger, H., Trelle, S., & Barth, J. (2011). Testing the allegiance bias hypothesis: A meta-analysis. *Psychotherapy Research*, *21*(6), 670–684. https://doi.org/10.1080/10503307.2011.602752

†Munder, T., Karcher, A., Yadikar, Ö., Szeles, T., & Gumz, A. (2019). Focusing on patients' existing resources and strengths in cognitive-behavioral therapy and psychodynamic therapy: A systematic review and meta-analysis. *Zeitschrift Für Psychosomatische Medizin Und Psychotherapie.*, *65*(2), 144–161. https://doi.org/10.13109/zptm.2019.65.2.144

*Muran, J. C., Safran, J. D., Samstag, L. W., & Winston, A. (2005). Evaluating an alliance-focused treatment for personality disorders. *Psychotherapy.*, *42*(4), 532–545. https://doi.org/10.1037/0033-3204.42.4.532

*Murphy, G. E., Carney, R. M., Knesevich, M. A., Wetzel, R. D., & Whitworth, P. (1995). Cognitive behavior therapy, relaxation training, and tricyclic antidepressant medication in the treatment of depression. *Psychological Reports*, *77*(2), 403–420. https://doi.org/10.2466/pr0.1995.77.2.403

National Health Service (2016). *Adult Improving Access to Psychological Therapies programme*. https://web.archive.org/web/20180129220631/england.nhs.uk/mental-health/adults/iapt/

†Newton-Howes, G., & Wood, R. (2013). Cognitive behavioural therapy and the psychopathology of schizophrenia: Systematic review and meta-analysis. *Psychology and Psychotherapy : Theory, Research and Practice.*, *86*(2), 127–138. https://doi.org/10.1111/j.2044-8341.2011.02048.x

*Nijdam, M. J., Gersons, B. P. R., Reitsma, J. B., de Jongh, A., & Olff, M. (2012). Brief eclectic psychotherapy v. eye movement desensitisation and reprocessing therapy for post-

traumatic stress disorder: randomised controlled trial. *The British Journal of Psychiatry.*, *200*(3), 224–231. https://doi.org/10.1192/bjp.bp.111.099234

*Nixon, R. D. (2012). Cognitive processing therapy versus supportive counseling for acute stress disorder following assault: A randomized pilot trial. *Behavior Therapy.*, *43*(4), 825–836. https://doi.org/10.1016/j.beth.2012.05.001

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). Jossey-Bass.

Norcross, J. C. (2005). A primer on psychotherapy integration. In J. C. Norcross, & M. R. Goldfried (Eds.) *Handbook of psychotherapy integration* (2nd ed., pp. 3–23). Oxford University Press.

Norcross, J. C., & Lambert, M. J. (Eds.). (2019). *Psychotherapy relationships that work. Volume 1: Evidence-based therapist contributions* (3rd ed.). Oxford University Press.

Norcross, J. C., & Wampold, B. E. (Eds.). (2019). *Psychotherapy relationships that work. Volume 2: Evidence-based therapist responsiveness* (3rd ed.). Oxford University Press.

*Nordahl, H. M., Borkovec, T. D., Hagen, R., Kennair, L. E. O., Hjemdal, O., Solem, S., Hansen, B., Haseth, S., & Wells, A. (2018). Metacognitive therapy versus cognitive–behavioural therapy in adults with generalised anxiety disorder. *BJPsych Open.*, *4*(5), 393–400. https://doi.org/10.1192/bjo.2018.54

†Normann, N., van Emmerik, A. A. P., & Morina, N. (2014). The efficacy of metacognitive therapy for anxiety and depression: A meta-analytic review. *Depression and Anxiety.*, *31*(5), 402–411. https://doi.org/10.1002/da.22273

Nunberg, H., & Federn, E. (Eds.). (1974). *Minutes of the Vienna Psychoanalytic Society, volume III: 1910-1911* (M. Nunberg & H. Collins, Trans.). International Universities Press.

*Øktedalen, T., Hoffart, A., & Langkaas, T. F. (2015). Trauma-related shame and guilt as time-varying predictors of posttraumatic stress disorder symptoms during imagery exposure and imagery rescripting—A randomized controlled trial. *Psychotherapy Research*, *25*(5), 518–532. https://doi.org/10.1080/10503307.2014.917217

OpenAI. (2026). *ChatGPT* (January 7 version) [Large language model]. https://chatgpt.com

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics.*, *8*(2), 157–159. https://doi.org/10.3102/10769986008002157

*Öst, L.-G. (1988). Applied relaxation vs progressive relaxation in the treatment of panic disorder. *Behaviour Research and Therapy.*, *26*(1), 13–22. https://doi.org/10.1016/0005-7967(88)90029-0

*Öst, L.-G., & Westling, B. E. (1995). Applied relaxation vs cognitive behavior therapy in the treatment of panic disorder. *Behaviour Research and Therapy.*, *33*(2), 145–158. https://doi.org/10.1016/0005-7967(94)E0026-F

*Öst, L.-G., Alm, T., Brandberg, M., & Breitholtz, E. (2001). One vs five sessions of exposure and five sessions of cognitive therapy in the treatment of claustrophobia. *Behaviour Research and Therapy.*, *39*(2), 167–183. https://doi.org/10.1016/S0005-7967(99)00176-X

*Öst, L.-G., Fellenius, J., & Sterner, U. (1991). Applied tension, exposure in vivo, and tension-only in the treatment of blood phobia. *Behaviour Research and Therapy.*, *29*(6), 561–574. https://doi.org/10.1016/0005-7967(91)90006-O

†Ougrin, D. (2011). Efficacy of exposure versus cognitive therapy in anxiety disorders: systematic review and meta-analysis. *BMC Psychiatry*, *11*(1), Article 200. https://doi.org/10.1186/1471-244X-11-200

†Papola, D., Miguel, C., Mazzaglia, M., Franco, P., Tedeschi, F., Romero, S. A., Patel, A. R., Ostuzzi, G., Gastaldon, C., Karyotaki, E., Harrer, M., Purgato, M., Sijbrandij, M., Patel, V., Furukawa, T. A., Cuijpers, P., & Barbui, C. (2024). Psychotherapies for generalized anxiety disorder in adults. *JAMA Psychiatry.*, *81*(3). https://doi.org/10.1001/jamapsychiatry.2023.3971

†Papola, D., Ostuzzi, G., Tedeschi, F., Gastaldon, C., Purgato, M., Del Giovane, C., Pompoli, A., Pauley, D., Karyotaki, E., Sijbrandij, M., Furukawa, T. A., Cuijpers, P., & Barbui, C. (2022). Comparative efficacy and acceptability of psychotherapies for panic disorder with or without agoraphobia: Systematic review and network meta-analysis of randomised controlled trials. *The British Journal of Psychiatry.*, *221*(3), 507–519. https://doi.org/10.1192/bjp.2021.148

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, *31*(2), 109–118. https://doi.org/10.1037/h0024436

*Paunovic, N., & Öst, L.-G. (2001). Cognitive-behavior therapy vs exposure therapy in the treatment of PTSD in refugees. *Behaviour Research and Therapy.*, *39*(10), 1183–1197. https://doi.org/10.1016/S0005-7967(00)00093-0

†Podina, I. R., Vîslă, A., Fodor, L. A., & Flückiger, C. (2019). Is there a sleeper effect of exposure-based vs. cognitive-only intervention for anxiety disorders? A longitudinal multilevel meta-analysis. *Clinical Psychology Review.*, *73*. https://doi.org/10.1016/j.cpr.2019.101774

*Power, K., McGoldrick, T., Brown, K., Buchanan, R., Sharp, D., Swanson, V., & Karatzias, A. (2002). A controlled comparison of eye movement desensitization and reprocessing versus exposure plus cognitive restructuring versus waiting list in the treatment of post-

traumatic stress disorder. *Clinical Psychology & Psychotherapy.*, *9*(5), 299–318. https://doi.org/10.1002/cpp.341

Prochaska, J. O., & DiClemente, C. C. (1982). Transtheoretical therapy: Toward a more integrative model of change. *Psychotherapy*, *19*(3), 276–288. https://doi.org/10.1037/h0088437

Rachman, S. (1973). The effects of psychological treatment. In H. J. Eysenck (Ed.), *Handbook of abnormal psychology* (pp. 805–861). Pitman.

Rachman, S., & Wilson, G. T. (1980). *The effects of psychological therapy*. Pergamon Press.

*Reger, G. M., Koenen-Woods, P., Zetocha, K., Smolenski, D. J., Holloway, K. M., Rothbaum, B. O., Difede, J., Rizzo, A. A., Edwards-Stewart, A., Skopp, N. A., Mishkind, M., Reger, M. A., & Gahm, G. A. (2016). Randomized controlled trial of prolonged exposure using imaginal exposure vs. virtual reality exposure in active duty soldiers with deployment-related posttraumatic stress disorder (PTSD). *Journal of Consulting and Clinical Psychology.*, *84*(11), 946–959. https://doi.org/10.1037/ccp0000134

*Resick, P. A., Nishith, P., Weaver, T. L., Astin, M. C., & Feuer, C. A. (2002). A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. *Journal of Consulting and Clinical Psychology.*, *70*(4), 867–879. https://doi.org/10.1037/0022-006X.70.4.867

*Richards, D. A., Ekers, D., McMillan, D., Taylor, R. S., Byford, S., Warren, F. C., Barrett, B., Farrand, P. A., Gilbody, S., Kuyken, W., O'Mahen, H., Watkins, E. R., Wright, K. A., Hollon, S. D., Reed, N., Rhodes, S., Fletcher, E., & Finning, K. (2016). Cost and Outcome of Behavioural Activation versus Cognitive Behavioural Therapy for

Depression (COBRA): a randomised, controlled, non-inferiority trial. *The Lancet.*, *388*(10047), 871–880. https://doi.org/10.1016/S0140-6736(16)31140-0

Riley, C., Lee, M., Cooper, Z., Fairburn, C. G., & Shafran, R. (2007). A randomised controlled trial of cognitive-behaviour therapy for clinical perfectionism: A preliminary study. *Behaviour Research and Therapy*, *45*(9), 2221–2231. https://doi.org/10.1016/j.brat.2006.12.003

Robinson, L. A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of controlled outcome research. *Psychological Bulletin, 108*(1), 30–49. https://doi.org/10.1037/0033-2909.108.1.30

Rogers, C. R. (1965). *Client-centered therapy: Its current practice, implications, and theory*. Houghton Mifflin.

Rosenfeld, E. A., & McLean, C. P. (2023). Purple hat therapies. In J. N. Stea & S. Hupp (Eds.), *Investigating clinical psychology: Pseudoscience, fringe science, and controversies* (pp. 116–126). Routledge.

Rosenthal, R. (1991). Meta-analysis: a review. *Psychosomatic Medicine*, *53*(3), 247–271. https://doi.org/10.1097/00006842-199105000-00001

Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *The American Journal of Orthopsychiatry, 6*(3), 412–415. https://doi.org/10.1111/j.1939-0025.1936.tb05248.x

Rosenzweig, S. (1985). Freud and experimental psychology: The emergence of idiodynamics. In S. Koch & D. E. Leary (Eds.), *A century of psychology as science* (pp. 135–207). McGraw-Hill.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology*, *58*(5), 646–656. https://doi.org/10.1037/0022-006X.58.5.646

†Ruiz, F. J. (2012). Acceptance and commitment therapy versus traditional cognitive behavioral therapy: A systematic review and meta-analysis of current empirical evidence. *International Journal of Psychology and Psychological Therapy*, *12*(3), 333–358.

Sanders, S. G., & Hunsley, J. (2018). The new caucus-race: Methodological considerations for meta-analyses of psychotherapy outcome. *Canadian Psychology, 59*(4), 387–398. https://doi.org/10.1037/cap0000164

*Sannibale, C., Teesson, M., Creamer, M., Sitharthan, T., Bryant, R. A., Sutherland, K., Taylor, K., Bostock-Matusko, D., Visser, A., & Peek-O'Leary, M. (2013). Randomized controlled trial of cognitive behaviour therapy for comorbid post-traumatic stress disorder and alcohol use disorders. *Addiction.*, *108*(8), 1397–1410. https://doi.org/10.1111/add.12167

*Scholing, A., & Emmelkamp, P. M. (1993). Cognitive and behavioural treatments of fear of blushing, sweating or trembling. *Behaviour Research and Therapy*, *31*(2), 155–170. https://doi.org/10.1016/0005-7967(93)90067-5

*Schramm, E., Kriston, L., Zobel, I., Bailer, J., Wambach, K., Backenstrass, M., Klein, J. P., Schoepf, D., Schnell, K., Gumz, A., Bausch, P., Fangmeier, T., Meister, R., Berger, M., Hautzinger, M., & Härter, M. (2017). Effect of disorder-specific vs nonspecific psychotherapy for chronic depression. *JAMA Psychiatry*, *74*(3), 233–242. https://doi.org/10.1001/jamapsychiatry.2016.3880

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMJ, 340*(7748), Article c332. https://doi.org/10.1136/bmj.c332

*Shalev, A. Y., Ankri, Y., Israeli-Shalev, Y., Peleg, T., Adessky, R., & Freedman, S. (2012). Prevention of posttraumatic stress disorder by early treatment: Results from the Jerusalem Trauma Outreach and Prevention study. *Archives of General Psychiatry*, *69*(2), 166–176. https://doi.org/10.1001/archgenpsychiatry.2011.127

Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin*, *92*(3), 581–604. https://doi.org/10.1037/0033-2909.92.3.581

Shapiro, F. (2001). *Eye movement desensitization and reprocessing (EMDR) therapy: Basic principles, protocols, and procedures* (2nd ed.). Guilford Press.

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., ... & 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*(8122), Article j4008. https://doi.org/10.1136/bmj.j4008

*Shear, M. K., Pilkonis, P. A., Cloitre, M., & Leon, A. C. (1994). Cognitive behavioral treatment compared with nonprescriptive treatment of panic disorder. *Archives of General Psychiatry*, *51*(5), 395–401. https://doi.org/10.1001/archpsyc.1994.03950050055006

†Siev, J., & Chambless, D. L. (2007). Specificity of treatment effects: Cognitive therapy and relaxation for generalized anxiety and panic disorders. *Journal of Consulting and Clinical Psychology*, *75*(4), 513–522. https://doi.org/10.1037/0022-006X.75.4.513

Sifneos, P. (1972). *Short-term psychotherapy and emotional crisis*. Harvard University Press.

*Simpson, H. B., Foa, E. B., Liebowitz, M. R., Ledley, D. R., Huppert, J. D., Cahill, S., Vermes, D., Schmidt, A. B., Hembree, E., Franklin, M., Campeas, R., Hahn, C.-G., & Petkova, E. (2008). A randomized, controlled trial of cognitive-behavioral therapy for augmenting pharmacotherapy in obsessive-compulsive disorder. *The American Journal of Psychiatry.*, *165*(5), 621–630. https://doi.org/10.1176/appi.ajp.2007.07091440

*Simpson, H. B., Zuckoff, A. M., Maher, M. J., Page, J. R., Franklin, M. E., Foa, E. B., Schmidt, A. B., & Wang, Y. (2010). Challenges using motivational interviewing as an adjunct to exposure therapy for obsessive-compulsive disorder. *Behaviour Research and Therapy.*, *48*(10), 941–948. https://doi.org/10.1016/j.brat.2010.05.026

†Singh, S. K., & Gorey, K. M. (2018). Relative effectiveness of mindfulness and cognitive behavioral interventions for anxiety disorders: Meta-analytic review. *Social Work in Mental Health.*, *16*(2), 238–251. https://doi.org/10.1080/15332985.2017.1373266

*Sloan, D. M., Marx, B. P., Lee, D. J., & Resick, P. A. (2018). A brief exposure-based treatment vs cognitive processing therapy for posttraumatic stress disorder: A randomized noninferiority clinical trial. *JAMA Psychiatry.*, *75*(3), 233–239. https://doi.org/10.1001/jamapsychiatry.2017.4249

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

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy*. Johns Hopkins University Press.

†Smith, M. M., & Hewitt, P. L. (2024). The equivalence of psychodynamic therapy and cognitive behavioral therapy for depressive disorders in adults: A meta-analytic review. *Journal of Clinical Psychology.*, *80*(5), 945–967. https://doi.org/10.1002/jclp.23649

*Solomonov, N., Falkenström, F., Gorman, B. S., McCarthy, K. S., Milrod, B., Rudden, M. G., Chambless, D. L., & Barber, J. P. (2020). Differential effects of alliance and techniques on panic-specific reflective function and misinterpretation of bodily sensations in two treatments for panic. *Psychotherapy Research*, *30*(1), 97–111. https://doi.org/10.1080/10503307.2019.1585591

†Spielmans, G. I., Benish, S. G., Marin, C., Bowman, W. M., Menster, M., & Wheeler, A. J. (2013). Specificity of psychological treatments for bulimia nervosa and binge eating disorder? A meta-analysis of direct comparisons. *Clinical Psychology Review.*, *33*(3), 460–469. https://doi.org/10.1016/j.cpr.2013.01.008

*Stangier, U., Schramm, E., Heidenreich, T., Berger, M., & Clark, D. M. (2011). Cognitive therapy vs interpersonal psychotherapy in social anxiety disorder. *Archives of General Psychiatry.*, *68*(7). https://doi.org/10.1001/archgenpsychiatry.2011.67

Statistics Canada. (2017, March 22). *Canadian community health survey, 2015* [Press release] https://www150.statcan.gc.ca/n1/en/daily-quotidien/170322/dq170322a-eng.pdf?st=aSt1mWPb

Stern, A., Skalsky, K., Avni, T., Carrara, E., Leibovici, L., & Paul, M. (2017). Corticosteroids for pneumonia. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD007720.pub3

Sterne, J.A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., … & 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

Strubell, E., Ganesh, A., & McCallum, A. (2020). Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI conference on artificial intelligence*

(AAAI-20), 34(9), 13693-13696. https://cdn.aaai.org/ojs/7123/7123-13-10352-1-10-20200526.pdf

Strubell, Ganesh, & McCallum, 2020

Strupp, H. H. (1963). The outcome problem in psychotherapy revisited. *Psychotherapy: Theory, Research & Practice, 1*(1), 1–13. https://doi.org/10.1037/h0088565

Svartberg, M., & Stiles, T. C. (1991). Comparative effects of short-term psychodynamic psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology, 59*(5), 704–714. https://doi.org/10.1037/0022-006x.59.5.704

*Svartberg, M., Stiles, T. C., & Seltzer, M. H. (2004). Randomized, controlled trial of the effectiveness of short-term dynamic psychotherapy and cognitive therapy for cluster c personality disorders. *The American Journal of Psychiatry.*, *161*(5), 810–817. https://doi.org/10.1176/appi.ajp.161.5.810

Swift, J. K., & Wampold, B. E. (2018). Inclusion and exclusion strategies for conducting meta-analyses. *Psychotherapy Research*, *28*(3), 356–366. https://doi.org/10.1080/10503307.2017.1405169

Tasca, G. A., Sylvestre, J., Balfour, L., Chyurlia, L., Evans, J., Fortin-Langelier, B., Francis, K., Gandhi, J., Huehn, L., Hunsley, J., Joyce, A. S., Kinley, J., Koszycki, D., Leszcz, M., Lybanon-Daigle, V., Mercer, D., Ogrodniczuk, J. S., Presniak, M., Ravitz, P., & Ritchie, K. (2015). What clinicians want: Findings from a psychotherapy practice research network survey. *Psychotherapy.*, *52*(1), 1–11. https://doi.org/10.1037/a0038252

*Taylor, S., Thordarson, D. S., Maxfield, L., Fedoroff, I. C., Lovell, K., & Ogrodniczuk, J. (2003). Comparative efficacy, speed, and adverse effects of three PTSD treatments:

Exposure therapy, EMDR, and relaxation training. *Journal of Consulting and Clinical Psychology.*, *71*(2), 330–338. https://doi.org/10.1037/0022-006X.71.2.330

Therapy deficit [Editorial]. 2012. *Nature, 489*, 473–474. https://doi.org/10.1038/489473b

Thoma, N., Pilecki, B., & McKay, D. (2015). Contemporary cognitive behavior therapy: A review of theory, history, and evidence. *Psychodynamic Psychiatry*, *43*(3), 423–461. https://doi.org/10.1521/pdps.2015.43.3.423

†Tolin, D. F. (2010). Is cognitive–behavioral therapy more effective than other therapies? A meta-analytic review. *Clinical Psychology Review*, *30*(6), 710–720. https://doi.org/10.1016/j.cpr.2010.05.003

†Tolin, D. F. (2014). Beating a dead dodo bird: Looking at signal vs. noise in cognitive-behavioral therapy for anxiety disorders. *Clinical Psychology: Science and Practice*, *21*(4), 351–362. https://doi.org/10.1111/cpsp.12080

Tolin, D. F. (2015), Corrigendum. *Clinical Psychology: Science and Practice, 22*(3), 315–316. https://doi.org/10.1111/cpsp.12109

†Tran, U. S., & Gregor, B. (2016). The relative efficacy of bona fide psychotherapies for post-traumatic stress disorder: A meta-analytical evaluation of randomized controlled trials. *BMC Psychiatry*, *16*(266), 1–21. https://doi.org/10.1186/s12888-016-0979-2

*Treasure, J., Todd, G., Brolly, M., Tiller, J., Nehmed, A., & Denman, F. (1995). A pilot study of a randomised trial of cognitive analytical therapy vs educational behavioral therapy for adult anorexia nervosa. *Behaviour Research and Therapy.*, *33*(4), 363–367. https://doi.org/10.1016/0005-7967(94)00070-Z

†Turner, D. T., Reijnders, M., van der Gaag, M., Karyotaki, E., Valmaggia, L. R., Moritz, S., Lecomte, T., Turkington, D., Penadés, R., Elkis, H., Cather, C., Shawyer, F., O'Connor,

K., Li, Z.-J., de Paiva Barretto, E. M., & Cuijpers, P. (2020). Efficacy and moderators of cognitive behavioural therapy for psychosis versus other psychological interventions: An individual-participant data meta-analysis. *Frontiers in Psychiatry*, *11*, 402. https://doi.org/10.3389/fpsyt.2020.00402

†Turner, D. T., van der Gaag, M., Karyotaki, E., & Cuijpers, P. (2014). Psychological interventions for psychosis: A meta-analysis of comparative outcome studies. *The American Journal of Psychiatry.*, *171*(5), 523–538. https://doi.org/10.1176/appi.ajp.2013.13081159

*Twohig, M. P., Abramowitz, J. S., Smith, B. M., Fabricant, L. E., Jacoby, R. J., Morrison, K. L., Bluett, E. J., Reuman, L., Blakey, S. M., & Ledermann, T. (2018). Adding acceptance and commitment therapy to exposure and response prevention for obsessive-compulsive disorder: A randomized controlled trial. *Behaviour Research and Therapy.*, *108*, 1–9. https://doi.org/10.1016/j.brat.2018.06.005

UK Council for Psychotherapy (n.d.). *New funded trainee places to be made available in the NHS*. https://www.psychotherapy.org.uk/news/new-funded-trainee-places-available-in-the-nhs/

*Vaccaro, L. D., Jones, M. K., Menzies, R. G., & Wootton, B. M. (2014). The treatment of obsessive-compulsive checking: A randomised trial comparing danger ideation reduction therapy with exposure and response prevention. *Clinical Psychologist*, *18*(2), 74–95. https://doi.org/10.1111/cp.12019

*van den Berg, D. P. G., de Bont, P. A. J. M., van der Vleugel, B. M., de Roos, C., de Jongh, A., Van Minnen, A., & van der Gaag, M. (2015). Prolonged exposure vs eye movement desensitization and reprocessing vs waiting list for posttraumatic stress disorder in

patients with a psychotic disorder: A randomized clinical trial. *JAMA Psychiatry, 72*(3), 259–267. https://doi.org/10.1001/jamapsychiatry.2014.2637

*van der Heiden, C., Muris, P., & van der Molen, H. T. (2012). Randomized controlled trial on the effectiveness of metacognitive therapy and intolerance-of-uncertainty therapy for generalized anxiety disorder. *Behaviour Research and Therapy.*, *50*(2), 100–109. https://doi.org/10.1016/j.brat.2011.12.005

*Verdellen, C. W., Keijsers, G. P., Cath, D. C., & Hoogduin, C. A. (2004). Exposure with response prevention versus habit reversal in Tourettes's syndrome: A controlled study. *Behaviour Research and Therapy.*, *42*(5), 501–511. https://doi.org/10.1016/S0005-7967(03)00154-2

*Vincelli, F., Anolli, L., Bouchard, S., Wiederhold, B. K., Zurloni, V., & Riva, G. (2003). Experiential cognitive therapy in the treatment of panic disorders with agoraphobia: A controlled study. *CyberPsychology and Behavior.*, *6*(3), 321–328. https://doi.org/10.1089/109493103322011632

*Visser, S., & Bouman, T. K. (2001). The treatment of hypochondriasis: Exposure plus response prevention vs cognitive therapy. *Behaviour Research and Therapy.*, *39*(4), 423–442. https://doi.org/10.1016/S0005-7967(00)00022-X

*Vogel, P. A., Stiles, T. C., & Götestam, K. G. (2004). Adding cognitive therapy elements to exposure therapy for obsessive compulsive disorder: A controlled study. *Behavioural and Cognitive Psychotherapy.*, *32*(3), 275–290. https://doi.org/10.1017/S1352465804001353

Von Ranson, K. M., Wallace, L. M., & Stevenson, A. (2013). Psychotherapies provided for eating disorders by community clinicians: Infrequent use of evidence-based

treatment. *Psychotherapy Research*, *23*(3), 333–343. https://doi.org/10.1080/10503307.2012.735377

*Vos, S. P. F., Huibers, M. J. H., Diels, L., & Arntz, A. (2012). A randomized clinical trial of cognitive behavioral therapy and interpersonal psychotherapy for panic disorder with agoraphobia. *Psychological Medicine.*, *42*(12), 2661–2672. https://doi.org/10.1017/S0033291712000876

Wampold, B. E. (2001). *The great psychotherapy debate: Models, methods, and findings* (1st ed.). Routledge.

Wampold, B. E., Budge, S. L., Laska, K. M., Del Re, A. C., Baardseth, T. P., Flűckiger, C., Minami, T., Kivlighan, D. M., & Gunn, W. (2011). Evidence-based treatments for depression and anxiety versus treatment-as-usual: A meta-analysis of direct comparisons. *Clinical Psychology Review*, *31*(8), 1304–1312. https://doi.org/10.1016/j.cpr.2011.07.012

Wampold, B. E., Flückiger, C., Del Re, A. C., Yulish, N. E., Frost, N. D., Pace, B. T., Goldberg, S. B., Miller, S. D., Baardseth, T. P., Laska, K. M., & Hilsenroth, M. J. (2017). In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy. *Psychotherapy Research*, *27*(1), 14–32. https://doi.org/10.1080/10503307.2016.1249433

Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate: The evidence for what makes psychotherapy work* (2nd ed.). Routledge.

Wampold, B. E., Imel, Z. E., & Miller, S. D. (2009). Barriers to the dissemination of empirically supported treatments: Matching messages to the evidence. *The Behaviour Therapist, 32*, 144–155.

†Wampold, B. E., Minami, T., Baskin, T. W., & Callen Tierney, S. (2002). A meta-(re)analysis of the effects of cognitive therapy versus 'other therapies' for depression. *Journal of Affective Disorders*, *68*(2), 159–165. https://doi.org/10.1016/S0165-0327(00)00287-1

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. N. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "all must have prizes." *Psychological Bulletin, 122*(3), 203–215. https://doi.org/10.1037/0033-2909.122.3.203

Wampold, B. E., & Serlin, R. C. (2014). Meta-analytic methods to test relative efficacy. *Quality & Quantity*, *48*(2), 755–765. https://doi.org/10.1007/s11135-012-9800-6

*Watson, J. C., Gordon, L. B., Stermac, L., Kalogerakos, F., & Steckley, P. (2003). Comparing the effectiveness of process-experiential with cognitive-behavioral psychotherapy in the treatment of depression. *Journal of Consulting and Clinical Psychology.*, *71*(4), 773–781. https://doi.org/10.1037/0022-006X.71.4.773

*Weck, F., Neng, J. M. B., Richtberg, S., Jakob, M., & Stangier, U. (2015). Cognitive therapy versus exposure therapy for hypochondriasis (health anxiety): A randomized controlled trial. *Journal of Consulting and Clinical Psychology.*, *83*(4), 665–676. https://doi.org/10.1037/ccp0000013

Weissman, M. M., Markowitz, J. C., & Klerman, G. L. (2000). *Comprehensive guide to Interpersonal Psychotherapy*. Basic Books.

Weisz, J. R., Jensen-Doss, A., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus usual clinical care: A meta-analysis of direct comparisons. *American Psychologist, 61*(7), 671–689. https://doi.org/10.1037/0003-066x.61.7.671

*Wells, A., Walton, D., Lovell, K., & Proctor, D. (2015). Metacognitive therapy versus prolonged exposure in adults with chronic post-traumatic stress disorder: A parallel randomized controlled trial. *Cognitive Therapy and Research.*, *39*(1), 70–80. https://doi.org/10.1007/s10608-014-9636-6

†Wen, H., Xiang, X., Jiang, Y., Zhang, H., Zhang, P., Chen, R., Wei, X., Dong, Y., Xiao, S., & Lu, L. (2023). Comparative efficacy of psychosocial interventions for opioid-dependent people receiving methadone maintenance treatment: A network meta-analysis. *Addiction.*, *118*(6), 1029–1039. https://doi.org/10.1111/add.16167

*Westra, H. A., Constantino, M. J., & Antony, M. M. (2016). Integrating motivational interviewing with cognitive-behavioral therapy for severe generalized anxiety disorder: An allegiance-controlled randomized clinical trial. *Journal of Consulting and Clinical Psychology.*, *84*(9), 768–782. https://doi.org/10.1037/ccp0000098

*Wetherell, J. L., Liu, L., Patterson, T. L., Afari, N., Ayers, C. R., Thorp, S. R., Stoddard, J. A., Ruberg, J., Kraft, A., Sorrell, J. T., & Petkus, A. J. (2011). Acceptance and commitment therapy for generalized anxiety disorder in older adults: A preliminary report. *Behavior Therapy.*, *42*(1), 127–134. https://doi.org/10.1016/j.beth.2010.07.002

*Whittal, M. L., Robichaud, M., Thordarson, D. S., & McLean, P. D. (2008). Group and individual treatment of obsessive-compulsive disorder using cognitive therapy and exposure plus response prevention: A 2-year follow-up of two randomized trials. *Journal of Consulting and Clinical Psychology.*, *76*(6), 1003–1014. https://doi.org/10.1037/a0013076

*Whittal, M. L., Woody, S. R., McLean, P. D., Rachman, S., & Robichaud, M. (2010). Treatment of obsessions: A randomized controlled trial. *Behaviour Research and Therapy.*, *48*(4), 295–303. https://doi.org/10.1016/j.brat.2009.11.010

*Wilhelm, S., Deckersbach, T., Coffey, B. J., Bohne, A., Peterson, A. L., & Baer, L. (2003). Habit reversal versus supportive psychotherapy for Tourette's disorder: A randomized controlled trial. *The American Journal of Psychiatry.*, *160*(6), 1175–1177. https://doi.org/10.1176/appi.ajp.160.6.1175

Wilson, G. T., & Rachman, S. J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. *Journal of Consulting and Clinical Psychology, 51*(1), 54–64. https://doi.org/10.1037/0022-006x.51.1.54

*Winston, A., Laikin, M., Pollack, J., Samstag, L. W., McCullough, L., & Muran, J. C. (1994). Short-term psychotherapy of personality disorders. *The American Journal of Psychiatry.*, *151*(2), 190–194. https://doi.org/10.1176/ajp.151.2.190

*Wolitzky, K. B., & Telch, M. J. (2009). Augmenting in vivo exposure with fear antagonistic actions: A preliminary test. *Behavior Therapy*, *40*(1), 57–71. https://doi.org/10.1016/j.beth.2007.12.006

Wolpe, J. (1948). *An approach to the problem of neurosis based on the conditioned response* [Unpublished M.D. thesis]. University of the Witwatersrand.

Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition*. Stanford University Press.

World Health Organization. (1979). *The ICD-9 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*.

Young, J. E., & Beck, A. T. (1980). Cognitive Therapy Rating Scale. Beck Institute for Cognitive Behavior Therapy. https://tinyurl.com/CTRSBeckYoung1980

Yulish, N. E., Goldberg, S. B., Frost, N. D., Abbas, M., Oleen-Junk, N. A., Kring, M., ... & Wampold, B. E. (2017). The importance of problem-focused treatments: A meta-analysis of anxiety treatments. *Psychotherapy, 54*(4), 321–338. https://doi.org/10.1037/pst0000144

*Zane, G., & Williams, S. L. (1993). Performance-related anxiety in agoraphobia: Treatment procedures and cognitive mechanisms of change. *Behavior Therapy.*, *24*(4), 625–643. https://doi.org/10.1016/S0005-7894(05)80322-3

*Zargar, F., Farid, A. A. A., Atef-Vahid, M.-K., Afshar, H., & Omidi, A. (2013). Comparing the effectiveness of acceptance-based behavior therapy and applied relaxation on acceptance of internal experiences, engagement in valued actions and quality of life in generalized anxiety disorder. *Journal of Research in Medical Sciences*, *18*(2), 118–122.

Zinbarg, R. E., Craske, M. G., & Barlow, D. H. (2006). *Mastery of your anxiety and worry: Therapist guide* (2nd ed.). Oxford University Press.

*Zuroff, D. C., Koestner, R., Moskowitz, D. S., McBride, C., Marshall, M., & Bagby, M. R. (2007). Autonomous motivation for therapy: A new common factor in brief treatments for depression. *Psychotherapy Research*, *17*(2), 137–147. https://doi.org/10.1080/10503300600919380

Appendix A: Search Terms for Table 1

The following terms were used to search the PsycINFO database in August 2021 for meta-analyses.

> ((psychodynamic or dynamic or "cognitive behavioural" or "cognitive-behavioural" or "cognitive behavioral" or "cognitive-behavioral" or "cognitive behavior therapy" or "cognitive-behavior therapy" or "cognitive behaviour therapy" or "cognitive-behaviour therapy" or CBT) and (psychotherap* or therap* or treat*) and (meta-analysis or meta analy* or metaanalysis or meta-analytic or metaanalytic or systematic)).ti,ab. and ("bona fide" or "direct comparison" or "direct comparisons" or Dodo or versus or prizes or relative).ti,ab

Collected studies were excluded if they did not include a meta-analysis examining direct comparisons of at least two different psychotherapies for individuals in adults.

Appendix B: Search Terms for Past Meta-Analyses on Relative Psychotherapy Efficacy

The following terms were used in August 2021 to search the PsycINFO database:

(psychotherap* or "CBT" or "IPT" or "PDT" or "EFT" or psychoanalysis or (psych* adj therap*) or (psych* adj interven*) or (psych* adj treat*) or (cognitive* adj therap*) or (cognitive* adj interven*) or (cognitive* adj treat*) or (behavio* adj therap*) or (behavio* adj interven*) or (behavio* adj treat*) or (interpersonal adj therap*) or (interpersonal adj interven*) or (interpersonal adj treat*) or (psychoanal* adj therap*) or (psychoanal* adj interven*) or (psychoanal* adj treat*) or (psychodynamic adj therap*) or (psychodynamic adj interven*) or (psychodynamic adj treat*) or (humanistic adj therap*) or (humanistic adj interven*) or (humanistic adj treat*) or (existential adj therap*) or (existential adj interven*) or (existential adj treat*) or "emotion-focused" or "emotionally-focused" or "dodo bird verdict").ti,ab.

(meta-analy* or (meta adj analy*) or metaanaly* or "systematic review" or "systematically review" or "systematically reviewed").ti,ab.

1 and 2

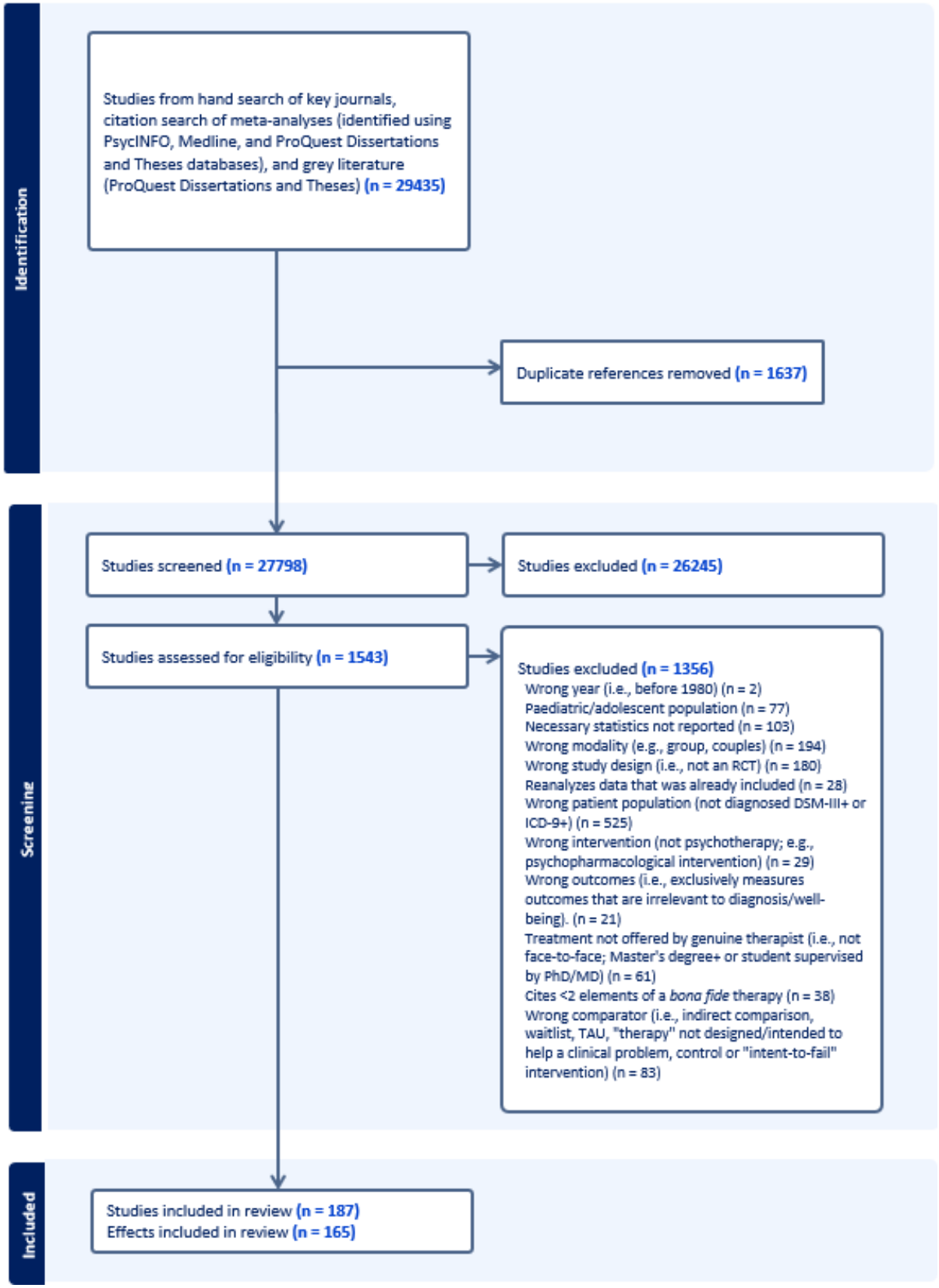
Limit 3 to (peer reviewed journal, full text)

The following terms were used to search the ProQuest Dissertations and Theses Global database:

(AB,TI(psychotherap*) OR AB,TI("CBT") OR AB,TI("IPT") OR AB,TI("PDT") OR AB,TI("EFT") OR AB,TI(psychoanalysis) OR AB,TI(psych*-therap*) OR AB,TI(psych*-interven*) OR AB,TI(psych*-treat*) OR AB,TI(cognitive*-therap*) OR AB,TI(cognitive*-interven*) OR AB,TI(cognitive*-treat*) OR AB,TI(behavio*-therap*) OR AB,TI(behavio*-interven*) OR AB,TI(behavio*-treat*) OR AB,TI(interpersonal-therap*) OR AB,TI(interpersonal-interven*) OR AB,TI(interpersonal-treat*) OR AB,TI(psychoanal*-therap*) OR AB,TI(psychoanal*-interven*) OR AB,TI(psychoanal*-treat*) OR AB,TI(psychodynamic-therap*) OR AB,TI(psychodynamic-interven*) OR AB,TI(psychodynamic-treat*) OR AB,TI(humanistic-therap*) OR AB,TI(humanistic-interven*) OR AB,TI(humanistic-treat*) OR AB,TI(existential-therap*) OR AB,TI(existential-interven*) OR AB,TI(existential-treat*) OR AB,TI("emotion-focused") OR AB,TI("emotionally-focused") OR AB,TI("dodo bird verdict"))

AND (AB,TI(meta-analysis) OR AB,TI(meta-analytic) OR AB,TI(metaanalysis) OR AB,TI(metaanalytic) OR AB,TI("systematic review") OR AB,TI("systematically review") OR AB,TI("systematically reviewed"))

Appendix C: PRISMA Flowchart



Appendix D: Risk of Bias 2 (RoB2) Scores for Primary Studies

Authors	Year of publication	ITT	RoB2 Randomiza-tion	RoB2 Assignm-ent	RoB2 Adheren-ce	RoB2 Missing Data	RoB2 Measure-ment	RoB2 Selection Bias	RoB2 Overall
Afshari, Hasani	2020	0	1	0	0	0	2	1	0
Agras, Schneider, Arnow, Raeburn, Telch	1989	0	1	0	0	0	2	0	0
Agras, Walsh, Fairburn, Wilson, Kraemer	2000	1	0	2	0	0	1	1	0
Arch, Eifert, Davies, Plumb Vilardaga, Rose, Craske	2012	1	0	2	2	0	2	1	0
Areán, Raue, Mackin, Kanellopoulos, McCulloch, Alexopoulos	2010	1	1	2	2	0	2	1	0
Arntz, Van Den Hout	1996	0	1	1	0	2	2	1	1
Babor, Carroll, Christiansen, Donaldson, Herrell, Kadden, Litt, McRee, Miller, Roffman, Solowji, Steinberg, Stephens, Vendetti	2004	0	2	0	0	2	2	2	0
Barber, Mildrod, Gallop, Solomonov, Rudden, McCarthy, Chambless	2020	0	0	0	0	0	2	0	0
Barlow, Craske, Cerny, Klosko	1989	0	1	0	0	0	2	1	0
Barlow, Farchione, Bullis, Gallagher, Murray-Latin, Sauer-Zavala, Bentley, Thompson-Hollands, Conklin, Boswell, Ametaj, Carl, Boettcher, Cassiello-Robbins	2017	1	2	2	2	0	2	0	0
Barlow, Rapee, Brown	1992	0	0	0	0	0	2	1	0
Barrowclough, Colville, Russell, Burns, Tarrier	2001	0	2	0	0	0	2	1	0
Belanger, Harvey, Fortier-Brochu, Beaulieu-Bonneau, Eidelman, Talbot, Ivers, Hein, Lamy, Soehner, Mérette, Morin	2016	1	2	2	2	2	2	1	1
Bellino, Zizza, Rinaldi, Bogetto	2007	0	1	0	0	0	2	1	0
Belloch, Cabello, Carrio	2008	0	1	0	0	0	2	1	0
Bernecker, Constantino, Atkinson, Bagby, Ravitz, and McBride	2016	0	0	0	0	0	2	1	0
Beutel, Scheurich, Knebel, Michal, Wiltink, Graf-Morgenstern, Tschan, Milrod, Wellek, Subic-Wrana	2013	0	1	0	0	0	2	0	0
Beutler, Moleiro, Malik, Harwood, Romanelli, Gallagher-Thompson, Thompson	2003	0	0	0	0	0	2	1	0
Blakey, Abramowitz, Buchholz, Jessup, Jacoby, Reuman, Pentel	2019	1	2	2	2	2	2	2	2
Blanco, Markowitz, Hellerstein, Nezu, Wall, Olfson, Chen, Levenson, Onishi, Varona, Okuda, Hershman	2019	0	1	2	0	0	2	1	0
Bodenmann, Plancherel, Beach, Widmer, Gabriel, Meuwly, Charvoz, Hautzinger, Schramm	2008	0	1	0	2	0	2	1	0
Borkovec, Costello	1993	0	1	0	0	0	2	1	0
Borkovec, Mathews	1988	0	2	0	0	0	2	1	0
Borkovec, Newman, Pincus, Lytle	2002	0	1	0	0	2	0	1	0
Bornas, Tortella-Feliu, Llabrés, Fullana	2001	0	0	0	0	0	2	1	0

Authors	Year of publication	ITT	RoB2 Randomiza-tion	RoB2 Assignm-ent	RoB2 Adheren-ce	RoB2 Missing Data	RoB2 Measure-ment	RoB2 Selection Bias	RoB2 Overall
Bramoweth, Lederer, Youk, Germain, Chinman	2020	1	1	2	0	0	2	2	0
Brom, Kleber, Defares	1989	0	0	0	0	0	2	1	0
Bruijniks, Lemmens, Hollon, Peeters, Cuijpers, Arntz, Dingemanse, Willems, van Oppen, Twisk, van den Boogaard, Spijker, Bosmans, Huibers	2020	1	1	2	0	0	0	2	0
Bruijniks, Lemmens, Hollon, Peeters, Cuijpers, Arntz, Dingemanse, Willems, van Oppen, Twisk, van den Boogaard, Spijker, Bosmans, Huibers	2020	1	1	2	0	0	0	2	0
Bryant, Mastrodomenico, Felmingham, Hopwood, Kenny, Kandris, Caholl, Creamer	2008	1	2	2	2	0	2	1	0
Bryant, Mastrodomenico, Hopwood, Kenny, Cahill, Kandris, Taylor	2013	1	2	2	0	0	2	2	0
Bryant, Moulds, Guthrie, Dang, Nixon	2003	1	1	2	2	0	2	1	0
Bryant, Moulds, Guthrie, Nixon	2005	1							0
Bryant, Sackville, Dang, Moulds, Guthrie	1999	0	1	0	2	0	2	1	0
Buckner, Zvolensky, Ecker, Schmidt, Lewis, Paulus, Lopez-Gamundi, Crapanzano, Bakhshaie	2019	0	1	0	2	0	2	1	0
Budney, Higgins, Radonovich, Novy	2000	1	0	2	0	0	2	1	0
Budney, Moore, Rocha, Higgins	2006	1	1	2	0	0	2	1	0
Burke, Drummond, Johnson	1997	0	0	0	0	0	2	2	0
Burling, Snyder Burling, Latini	2001	0	1	2	2	0	2	0	0
Butler, Fennell, Robson, Gelder	1991	0	1	1	2	2	2	1	1
Butollo, Karl, Konig, Rosner	2016	1	0	0	0	2	2	1	0
Capezzani, Ostacoli, Cavallo, Carletto, Fernandez, Solomon, Pagani, Cantelmi	2013	1	1	2	0	2	2	0	0
Carter, Crowe, Carlyle, Frampton, Jordan, McIntosh, O'Toole, Whitehead, Joyce	2012	1	1	2	2	0	0	1	0
Carter, McIntosh, Jordan , Porter, Frampton, Joyce	2013	1	1	2	2	0	1	2	0
Clark, Ehlers, Hackmann, McManus, Fennell, Grey, Waddington, Wild	2006	1	1	2	2	0	2	1	0
Clark, Salkovskis, Hackmann, Middleton, Anastasiades, Gelder	1994	0	1	0	0	0	2	1	0
Clark, Salkovskis, Hackmann, Wells, Femmell, Ludgate, Ahmad, Richards, Gelder	1998	0	0	0	2	0	2	1	0
Connolly Gibbons, Gallop, Thompson, Luther, Crits-Christoph, Jacobs, Yin, Crits-Christoph	2006	1	1	2	2	0	2	0	0
Constantino, Marnell, Haile, Kanther-Sista, Wolman, Zappert, Arnow	2008	1	0	2	0	0	2	1	0
Cottraux, Note ,Yao, Lafont, Note, Mollarda, Bouvard, Sauteraud, Bourgeois, Dartigues	2001	0	1	0	0	0	2	1	0
Cottraux, Note, Yao, de Mey-Guillard, Bonasse, Djamoussian, Mollard, Note, Chen	2008	0	1	2	0	0	2	2	0

Authors	Year of publication	ITT	RoB2 Randomiza-tion	RoB2 Assignm-ent	RoB2 Adheren-ce	RoB2 Missing Data	RoB2 Measure-ment	RoB2 Selection Bias	RoB2 Overall
Craske, Maidenberg, Bystritsky	1995	0							0
Craske, Meuret, Ritz, Treanor, Dour, Rosenfield	2019	0	1	0	0	0	2	1	0
Craske, Niles, Burklund, Wolitzky-Taylor, Plumb Vilardaga, Arch, Saxbe, Lieberman	2014	1	1	2	0	2	2	1	1
Crits-Christoph, Siqueland, Blaine, Frank, Luborsky, Onken, Muenz, Thase, Weiss, Gastfriend, Woody, Barber, Butler, Daley, Salloum, Bishop, Najavits, Lis, Mercer, Griffin, Moras, Beck	1999	1	2	2	0	0	2	2	0
Daniel, Poulsen, Lunn	2016	0	2	0	2	0	0	1	0
de Ruiter, Rijken, Garssen, Kraaimaat	1989	0	1	0	0	0	2	1	0
Deckersbach, Rauch, Buhlmann, Wilhelm	2006	0	0	0	0	2	2	1	0
Dimidjian, Hollon, Dobson, Schmaling, Kohlenberg, Addis, Gallop, McGlinchey, Markley, Gollan, Atkins, Dunner, Jacobson	2006	0	0	2	0	2	2	1	0
Driessen, Van, Don, Peen, Kool, Westra, Hendriksen, Schoevers, Cuijpers, Twisk, Dekker	2013	1	1	2	0	0	2	2	0
Drummond, Glautier	1994	0	1	0	0	0	2	1	0
Dugas, Brillon, Savard, Turcotte, Gaudet, Ladouceur, Leblanc, Gervais	2010	1	1	2	2	0	2	1	0
Durham, Murphy, Allan, Richard, Treliving, Fenton	1994	0	1	0	0	0	2	1	0
Eddington, Silvia, Foxworth, Hoet, Kwapil	2015	0	1	2	2	0	2	2	0
Ehlers, Hackman, Grey, Wild, Liness, Albert, Deale, Stott, Clark	2014	1	2	2	0	0	2	2	0
Ellison, Greenberg, Goldman, Angus	2009	0	1	0	2	0	2	1	0
Emmeikamp, Visser, Hoekstra	1988	0	1	0	0	0	2	1	0
Emmelkamp, Beens	1991	0	1	0	0	0	2	1	0
Emmelkamp, Benner, Kuipers, Feiertag, Koster, van Apeldoorn	2006	0	0	0	0	0	0	1	0
Emmelkamp, Krijn, Hulsbosch, de Vries, Schuemie, van der Mast	2002	0	1	1	0	0	2	1	0
Fairburn, Bailey-Straebler, Basden, Doll, Jones, Murphy, O'Connor, Cooper	2015	1	0	2	0	0	2	0	0
Farahimanesh, Moradi, Sadeghi, Jobson	2021	1	1	2	0	2	2	1	1
Flückiger, Forrer, Schnider, Bättig, Bodenmann, Zinbarg	2016	0	1	0	2	0	2	0	0
Foa, Dancu, Hembree, Jaycox, Meadows, Street	1999	0	1	0	2	0	2	1	0
Foa, Hembree, Cahill, Rauch, Riggs, Feeny, Yadin	2005	1	0	2	0	0	2	1	0
Foa, Rothbaum, Riggs, Murdock	1991	0	0	0	2	0	2	1	0
Fonagy, Lemma, Target, O'Keeffe, Constantinou, Ventura Wurman, Luyten, Allison, Roth, Cape, Pilling	2020	1	2	2	0	1	2	2	1
Ford, Steinberg, & Zhang	2011	1	1	0	0	0	2	1	0
Freeman, Barry, Dunkeld-Turnbull, Henderson	1988	0	0	0	0	0	2	1	0

Authors	Year of publication	ITT	RoB2 Randomiza-tion	RoB2 Assignm-ent	RoB2 Adheren-ce	RoB2 Missing Data	RoB2 Measure-ment	RoB2 Selection Bias	RoB2 Overall
Ghaderi	2006	1	1	2	2	2	2	1	1
Giesen-Bloo, Van Dyck, Spinhoven, Van Tilburg, Dirksen, Van Asselt,Kremers, Nadort, Arntz	2006	1	2	2	2	0	2	1	0
Gloster, Hauke, Höfler. Einsle, Fydrich, Hamm, Sthröhle, Wittchen	2013	1	1	2	0	0	2	0	0
Goldman, Greenberg, Angus	2006	0	1	0	2	0	2	1	0
Greenberg, Watson	1998	1	1	0	0	0	2	1	0
grosse Holtforth, Krieger, Zimmermann, Altenstein-Yamanaka, Dorig, Meisch, Hayes	2019	1	2	2	0	2	2	2	2
Hardy, Barkham, Shapiro, Stiles, Rees, Reynolds	1995	0	0	0	2	0	0	1	0
Harvey, Bélanger, Talbot, Eidelman, Beaulieu-Bonneau, Fortier-Brochu, Ivers, Lamy, Hein, Soehner, Mérette, Morin	2014	1	2	2	2	2	2	1	1
Hayes-Skelton, Roemer, Orsillo	2013	0	2	0	0	2	2	1	0
Hellstrom, Fellenius, Ost	1996	1	1	2	0	2	2	1	1
Hemanny, Carvalho ,Maia, Reis, Botelho, Bonavides, Seixas, de Oliveira	2020	0	1	0	2	0	2	2	0
Hersen, Bellack, Himmelhoch, Thase	1984	0	1	0	0	0	2	1	0
Hien, Cohen, Miele, Litt, Capstick	2004	1	2	2	2	1	2	1	1
Hien, Smith, Owens, Lopez-Castro, Ruglass, Papini	2018	0	2	0	2	0	2	1	0
Hoffart, Oktedalen, Langkaas, Wampold	2013	1	2	2	0	0	2	1	0
Hopko, Armento, Robertson, Ryba, Carvalho, Colman, Mullane, Gawrysiak, Bell, McNulty, Lejuez	2011	1	1	2	2	1	2	1	1
Horst, Den Oudsten, Zijlstra, de Jongh, Lobbestael, De Vries	2017	0	0	2	0	0	2	2	0
Hoyer, Beesdo, Gloster, Runge, Höfler, Becker	2009	0	1	1	2	0	2	1	0
Huber, Henrich, Gastner, Klug	2012	0	0	0	0	0	0	1	0
Ito, de Araujo, Tess, de Barros-Nero, Asbahr, Marks	2001	0	2	0	0	0	2	1	0
Jackson, Pietrabissa, Rossi, Manzoni, Castelnuovo	2018	1	2	2	2	2	2	0	0
Jacobson, Dobson, Truax, Addis, Koerner, Gollan, Gortner, Prince	1996	1	1	2	2	0	2	1	0
Janse, de Jong, Veerkamp, van Dijk, Hutschemaekers, Verbraak	2020	1	1	0	0	0	2	1	0
Kampmann, Emmelkamp, Hartanto, Brinkman, Zijlstra, Morina	2016	1	2	2	0	0	2	0	0
Keijsers, Maas, van Opdorp, van Minnen	2016	0	1	1	0	0	2	1	0
Kiosses, Arean, Teri, Alexopoulos	2010	1	1	2	0	0	2	1	0
Klein, Leon, Li, D'Zurilla, Black, Vivian, Dowling, Arnow, Manber, Markowitz, Kocsis	2011	0	2	0	2	0	2	2	0
Klingberg, Wolwer, Engel, Wittorf, Herrlich, Meisner, Buchkremer, Wiedemann	2011	1	1	0	0	0	2	1	0

Authors	Year of publication	ITT	RoB2 Randomiza-tion	RoB2 Assignm-ent	RoB2 Adheren-ce	RoB2 Missing Data	RoB2 Measure-ment	RoB2 Selection Bias	RoB2 Overall
Koszycki, Bisserbe, Blier, Bradwejn Markowitz	2012	1	1	2	0	0	2	1	0
Koszycki, Raab, Aldosary, Bradwejn	2010	1	1	2	0	0	2	1	0
Kramer, Kolly, Berthoud, Keller, Preisig, Caspar, Berger, de Roten, Marquet, Despland	2014	1	2	2	2	0	2	2	0
Kunze, Arntz, Morina, Kindt, Lancee	2017	1	2	2	2	0	2	2	0
Langkaas, Hoffart, Oktedalen, Ulvenes, Hembree, Smucker	2017	1	0	2	0	0	2	1	0
Leichsenring, Salzer, Beutel, Herpertz, Hiller, Hoyer, Huesing, Joraschky, Nolting, Poehlmann, Ritter, Stangier, Strauss, Stuhldreher, Tefikow, Teismann, Willutzki, Wiltink, Leibing	2013	1	1	2	2	2	2	1	1
Leichsenring, Salzer, Jaeger, Kachele, Kreische, Leweke, Ruger, Winkelbach, Leibing	2009	1	1	2	2	0	2	1	0
Lemmens, Arntz, Peeters, Hollon, Roefs, Huibers	2015	1	1	2	2	0	2	1	0
Li, Guo, Wang, Xu, Qu, Wang, Sun, Yan, Ng, Turkington, Kingdon	2015	1							0
Lipsitz, Gur, Vermes, Petkova, Cheng, Miller, Laino, Leibowitz, Fyer	2008	1	1	2	0	0	2	1	0
Lopes, Goncalves, Machado, Sinai, Bento, Salgado	2014	1	1	0	0	0	2	1	0
Maina, Rosso, Crespi, Bogetto	2007	1	1	0	0	0	2	1	0
Markowitz, Kocsis, Christos, Bleiberg, Carlin	2008	1	1	2	2	0	2	1	0
Markowitz, Petkova, Neria, Van Meter, Zhao, Hembree, Lovell, Biyanova, Marshall	2015	1	0	2	2	2	2	1	0
Marks, Lovell, Noshirvani, Livanou, Thrasher	1998	0	0	0	2	0	2	1	0
Martini, Rosso, Chiodelli, de Cori, Maina	2011	1	1	0	0	0	2	1	0
Marttunen, Valikoski, Lindfors, Laaksonen, Knekt	2008	0	2	0	0	0	2	1	0
McDonagh, Friedman, McHugo, Ford, Sengupta, Mueser, Demment, Fournier, Schnurr, Descamps	2005	0	1	0	2	0	2	1	0
McIndoo, File, Preddy, Clark, Hopko	2016	1	1	2	2	2	2	1	1
McKay, Lynch, Coviello, Morrison, Cary, Skalina, Plebani	2010	1	1	2	0	2	2	1	1
Mersch	1995	0	0	0	0	0	2	1	0
Meyer, Hautzinger	2012	1	1	0	0	0	2	1	0
Morgenstern, Irwin, Wainberg, Parsons, Muench, Bux, Kahler, Marcus, Schulz-Heik	2007	0	1	0	2	0	2	1	0
Muran, Safran, Wallner Samstag, Winston	2005	0	1	0	0	0	2	1	0
Murphy, Carney, Knesevich, Wetzel, Whitworth	1995	1	1	2	0	0	2	0	0
Nijdam, Gersons, Reitsma, de Jongh, Olff	2012	1	2	2	2	0	0	2	0
Nixon	2012	1	0	2	0	0	0	1	0
Nordahl, Borkovec, Hagen, Kennair, Hjemdal, Solem, Hansen, Haseth, Wells	2018	1	1	0	2	2	2	2	1

Authors	Year of publication	ITT	RoB2 Randomiza-tion	RoB2 Assignm-ent	RoB2 Adheren-ce	RoB2 Missing Data	RoB2 Measure-ment	RoB2 Selection Bias	RoB2 Overall
Øktedalen, Hoffart, Langkaas	2015	0	0	0	0	0	2	0	0
O'Malley, Jaffe, Chang, Schottenfeld, Meyer, Rounsaville	1992	1	1	0	0	0	2	1	0
Ost	1988	0	0	0	0	1	2	1	0
Ost & Westling	1995	0	0	0	0	1	2	1	0
Ost, Alm, Brandberg, Breitholtz	2001	0	1	0	0	0	0	1	0
Ost, Fellenius, Sterner	1991	1	1	1	0	2	2	1	1
Paunovic, Ost	2000	0	1	0	0	0	1	1	0
Power, McGoldrick, Brown, Buchanan, Sharp, Swanson, Karatzias	2002	0	1	0	0	0	2	1	0
Reger, Koenen-Woods, Zetocha, Smolenski, Holloway, Rothbaum, Difede, Rizzo, Edwards-Stewart, Skopp, Mishkind, Reger, Gahm	2016	1	1	2	2	0	2	1	0
Resick, Nishith, Waver, Asin, Feuer	2002	0	1	0	2	0	2	1	0
Richards,Ekers, McMillan, Taylor, Byford, Warren, Barrett, Farrand, Gilbody, Kuyken, O'Mahen, Watkins, Wright, Hollon, Reed, Rhodes, Fletcher, Finning	2016	1	2	2	0	0	2	2	0
Ritter, Leichsenring, Strauss, Stangier	2013	1	1	2	0	0	2	1	0
Sannibale, Teesson, Creamer, Sitharthan, Bryant, Sutherland, Taylor, Bostock-Matusko,Visser, Peek-O'Leary	2013	0	2	2	0	0	2	0	0
Schnurr, Friedman, Engel, Foa, Shea, Chow,Resick, Thurston, Orsillo, Haug, Turner, Bernardy	2007	1	1	0	0	0	2	1	0
Scholing, Emmelkamp	1993	0	1	0	0	0	2	1	0
Schramm, Kriston, Zobel, Bailer, Wambach, Backenstrass, Klein, Schoepf, Schnell, Gumz, Bausch, Fangmeier, Meister, Berger, Hautzinger, Harter	2017	1	1	2	0	2	2	2	1
Shalev, Ankri, Israeli-Shalev, Peleg, Adessky, Freedman	2012	0	1	0	2	0	0	0	0
Shear, Pilkonis, Cloitre, Leon	1994	0	0	0	0	0	2	1	0
Simpson, Foa, Liebowitz, Ledley, Huppert, Cahill, Vermes, Schmidt, Hembree, Franklin, Campeas, Hahn, Petkova	2008	0	1	0	2	0	2	1	0
Simpson, Zuckoff, Maher, Page, Franklin, Foa, Schmidt, Wang	2010	1	1	2	0	0	2	1	0
Sloan, Marx. Lee, Resick	2018	1	1	2	2	0	0	2	0
Solomonov, Falkenstrom, Gorman, McCarthy, Milrod, Rudden, Chambless, Barber	2019	0	1	0	0	0	2	2	0
Stangier, Schramm,Heidenreich, Berger, Clark	2011	1	2	2	2	0	2	1	0
Surís,Link-Malcolm, Chard, Ahn, North	2013	1							0
Svartberg, Stiles, Seltzer	2004	0	1	0	0	0	2	1	0
Taylor, Thordarson, Maxfield, Fedoroff, Lovell, Ogrodniczuk	2003	0	1	0	0	0	2	1	0
Treasure, Todd, Brolly, Tiller, Nehmed, Denman	1995	1	1	2	0	0	2	1	0

Authors	Year of publication	ITT	RoB2 Randomiza-tion	RoB2 Assignm-ent	RoB2 Adheren-ce	RoB2 Missing Data	RoB2 Measure-ment	RoB2 Selection Bias	RoB2 Overall
Turkington, Sensky, Scott, Barnes, Nur, Siddle, Hammond, Samarasekara, Kingdon	2008	1	1	2	2	0	2	1	0
Twohig, Abramowitz, Smith, Fabricant, Jacoby, Morrison, Bluett, Reuman, Blakey, Ledermann	2018	1	1	2	2	0	2	2	0
Twohig, Hayes, Plumb, Pruitt, Collins,Hazlett-Stevens, Woidneck	2010	1	1	0	0	0	2	1	0
Vaccaro, Jones, Menzies, Wootton	2014	1	1	2	2	0	2	1	0
Valmaggia, van der Gaag, Tarrier, Pijnenborg, Slooff	2005	1	1	0	2	0	2	1	0
van den Berg, de Bont, van der Vleugel, de Roos, de Jongh, Van Minnen, van der Gaag	2015	1	2	2	2	2	2	2	2
van der Heiden,Muris, van der Molen	2012	0	1	0	2	0	0	1	0
Verdellen, Keijsers, Cath, Hoogduin	2004	0	1	0	0	0	2	1	0
Vincelli, Anolli, Bouchard, Wiederhold, Zurloni, Riva	2003	1	0	2	0	2	2	2	0
Visser, Bouman	2001	0	0	0	0	0	2	1	0
Vogel, Stules, Gotestam	2004	1	2	2	2	0	0	1	0
Vos, Huibers, Diels, Arntz	2012	1	1	2	0	0	2	1	0
Watson, Gordon, Stermac,Kalogerakos, Steckley	2003	1	0	2	2	0	2	1	0
Weck, Nagel, Hofling, Neng	2017	0	1	0	2	0	0	1	0
Wells, Walton, Lovell, Proctor	2015	0	1	0	2	0	2	2	0
Westra, Constantino, Antony	2016	1	0	2	0	0	2	1	0
Wetherell, Afari, Ayers, Stoddard, Ruberg, Sorrell, Liu, Petkus, Thorp, Kraft, Patterson	2011	0	0	0	0	0	2	1	0
Whittal, Robichaud, Thordarson, McLean	2008	0	0	0	0	0	2	1	0
Whittal, Woody, McLean, Rachman, Robichaud	2010	0	1	0	0	0	0	1	0
Wilhelm, Deckersbach, Coffey, Bohne, Peterson, Baer	2003	0	0	0	0	0	2	1	0
Winston, Laikin, Pollack, Samstag, McCullough, Muran	1994	0	1	0	0	0	2	1	0
Wolitzky, Telch	2009	0	1	0	2	2	2	1	1
Zane, Williams	1993	0	1	0	2	0	0	1	0
Zargar, Farid, Atef-Vahid, Afshar, Omidi	2013	0	1	0	0	0	2	2	0
Zuroff, Koestner, Moskowitz, McBride, Marshall, Bagby	2007	0	1	0	0	0	0	1	0

Note. “ITT” = intent-to-treat analysis; 1 = yes, 0 = no (e.g., per-protocol analysis). Other numbers represent the score per the Risk of Bias 2 coding algorithms. Note scoring for the RoB2 domain 2 (“deviations from intended interventions”) was based on column 2 (“Assignment”) for studies using ITT analysis and based on column 3 (“Adherence”) for studies analyzed per protocol.

Appendix E: RoB2 Criteria

Risk of bias arising from the <u>randomization</u> process

- **Concealment?**
- **Baseline differences?**

2 points. The allocation sequence was concealed from enrolment personnel and participants until participants were enrolled and assigned to interventions. No indication of baseline differences between intervention groups that suggest problems with the randomization process.

1 point. The allocation sequence was concealed from enrolment personnel and participants until participants were enrolled and assigned to interventions, but baseline differences between intervention groups suggest problems with the randomization process. OR Insufficient information to determine if the allocation was concealed but no indication that baseline differences between intervention groups to suggest problems with the randomization process.

0 points. The allocation sequence was not concealed OR Insufficient information to determine if the allocation was concealed and there are baseline differences between intervention groups that suggest problems with the randomization process.

Score ONE of the following two items, depending on whether the study uses manages missing data using an ITT or per-protocol approach

Risk of bias due to the effect of <u>assignment</u> to intervention (ITT)

- **Deviations from protocol due to assignment?**
- **ITT?**

2 points.
Unlikely that deviations from intervention protocol arose due to trial context (i.e., effects of recruitment or engagement or due to people delivering interventions undermining the implementation of trial protocol in ways that would not happen outside the trial; e.g., process of informed consent leads to seeking “better” interventions), OR changes from assigned intervention were consistent with either trial protocol or with what would happen outside the trial context
AND
An appropriate analysis (ITT/mITT) was used to estimate the effect of assignment to intervention.

1 point. EITHER
No information about whether deviations from protocol arose from trial context OR there is indication that the trial context led to failure to implement the protocol interventions or to implementation of interventions not allowed by the protocol, but they were not likely to have affected outcome OR there were deviations that may have affected the outcome, but the deviations from intended intervention were balanced between groups
OR
There may not have been an appropriate analysis used to estimate the effect of assignment to

intervention, but the potential impact (on the result) of the failure to analyse participants in the group to which they were randomized was not substantial.

0 points. EITHER
Indication that the trial context led to deviations from the trial protocol and these deviations were likely to have affected the outcome, AND that they may not have been balanced between groups
OR
There may not have been an appropriate analysis used to estimate the effect of assignment to intervention, and the potential impact (on the result) of the failure to analyse participants in the group to which they were randomized was substantial.

Risk of bias due to deviations in <u>adherence</u> to the intended interventions (Per Protocol)

- **Balanced non-protocol interventions?**
- **Deviations from assigned interventions?**

2 points. Any important non-protocol interventions were balanced across intervention groups AND there were no failures in implementing the intervention, or in participant adherence (~~e.g., dropout)~~ to the assigned intervention regimen, that could have affected participants' outcomes.

1 point. One of the requirements in [2 points] was not clearly met, but there was an appropriate analysis used to estimate the effect of adhering to the intervention (e.g., instrumental variable analysis or inverse probability weighting, with justified assumptions and reporting of info on deviations from protocol)

0 points. One of the requirements in [2 points] was not clearly met and there was inadequate indication that an appropriate analysis was used to estimate the effect of adhering to the intervention.

Missing outcome <u>data</u>

- **95%+ available?**
- **Result not biased by missingness, per sensitivity analyses/bias corrections?**
- **Missingness not dependent on true value?**

2 points. There were data for this outcome available for nearly all (>95%) of randomized ITT participants, OR there is evidence that the result was not biased by missing outcome data (bias corrections, sensitivity analyses – not imputation or LOCF) OR missingness in the outcome could not depend on its true value (e.g., missing due to documented reasons unrelated to participant health, such as measurement device or data collection issues).

1 point. None of the requirements in [2 points] were met, AND unlikely that missingness in the outcome depended on its true value (e.g., no differences in proportion of missing data between groups; no reasons reported that connect missingness to its true value; no difference in reasons between groups for missingness; no reasons related to the trial circumstances that make it likely that missingness depends on true value [e.g., dropout related to symptom severity]).

0 points. None of the requirements in [2 points] were met, AND missingness in the outcome may have depended on its true value (e.g., differences in proportion of missing data between groups; reasons reported that connect missingness to its true value; different reasons between groups for missingness; likely reasons related to the trial circumstances that missingness depends on true value [e.g., dropout related to symptom severity]).

Risk of bias in measurement of the outcome

- **Appropriate measurement?**
- **No systematic differences in assessment?**

2 points. Method of measuring the outcome was not inappropriate AND measurement or ascertainment of the outcome probably could not have differed between intervention groups (i.e., no systematic differences in method, quantity of assessments)

1 point. Method of measuring the outcome was not inappropriate AND no information to determine whether measurement or ascertainment of the outcome could have differed between intervention groups.

0 points. Method of measuring the outcome was inappropriate (i.e., unlikely to be sensitive to plausible intervention effects [e.g. important ranges of outcome values fall outside levels that are detectable using the measurement method) or the measurement instrument has been demonstrated to have poor validity] OR measurement or ascertainment of the outcome could have differed between intervention groups.

Risk of bias in selection of the reported result

- **Effect not likely to have been cherrypicked?**
- **Pre-specified plan?**

2 points. There are indications that the numerical result being assessed is not likely to have been selected, on the basis of the results, from either (1) multiple eligible outcome measurements (e.g. scales, definitions, time points) within the outcome domain or (2) multiple eligible analyses of the data,

i.e., There is clear evidence (usually through examination of a trial protocol or statistical analysis plan) that all eligible reported results for the outcome domain correspond to all intended outcome measurements/analyses, OR There is only one possible way in which the outcome domain can be measured/analyzed (hence there is no opportunity to select from multiple measures/analyses), OR outcome measurements/analyses are inconsistent across different reports on the same trial, but the trialists have provided the reason for the inconsistency and it is not related to the nature of the results,

AND the data that produced this result were analysed in accordance with a pre-specified analysis plan that was finalized before unblinded outcome data were available for analysis.

1 point. No information about whether the numerical result being assessed is likely to have been selected, on the basis of the results, from either (1) multiple eligible outcome measurements (e.g. scales, definitions, time points) within the outcome domain or (2) multiple eligible analyses of the data.

i.e., Analysis intentions are not available, or the analysis intentions are not reported in sufficient detail to enable an assessment, and there is more than one way in which the outcome domain could have been measured/analyzed.

OR The result is not likely to have been selected from either multiple eligible outcome measurements or multiple analyses of the data, AND there isn't sufficient indication that the trial was analyzed in accordance with a pre-specified plan.

0 points. The numerical result being assessed is likely to have been selected, on the basis of the results, from multiple eligible outcome measurements (e.g. scales, definitions, time points) within the outcome domain OR multiple eligible analyses of the data.

i.e., There is clear evidence (usually through examination of a trial protocol or statistical analysis plan) that a domain was measured/analyzed in multiple eligible ways, but data for only one or a subset of measures/analyses is fully reported (without justification), and the fully reported result is likely to have been selected on the basis of the results. Selection on the basis of the results can arise from a desire for findings to be newsworthy, sufficiently noteworthy to merit publication, or to confirm a prior hypothesis. For example, trialists who have a preconception, or vested interest in showing, that an experimental intervention is beneficial may be inclined to report outcome measurements selectively that are favourable to the experimental intervention.

Appendix F: Adaptation of the Multitheoretical List of Therapeutic Interventions

(MULTI; McCarthy & Barber, 2009)

Item	Score 1 = not a match 2 = weak partial match 3 = strong partial match 4 = match
Therapist set an agenda or established specific goals for the therapy session	
Therapist made connections between client's current situation and past	
Therapist focused on identifying parts of client's personality that were in conflict, like one part that wanted to be close to others and another part that did not.	
therapist asked client to visualize specific scenes or situations in detail	
therapist encouraged client to identify specific situations or events that tended to precede client's problematic behavior.	
therapist often focused on client's recent experiences	
therapist and client discussed a plan for client to try to control (increase or decrease) specific behaviors (e.g., smoking, eating, exercising, checking, saying or thinking certain things, hurting themselves)	
therapist repeated back to client (paraphrased) the meaning of what they were saying	
therapist encouraged client to identify or label feelings that they had in or outside of the session	
therapist encouraged client to talk about feelings they had previously avoided or never expressed	
Client's therapist pointed out times when client's behavior seemed inconsistent with what they were saying (suddenly shifted moods/topics, long silences, laughed, smiled, looked away, was uncomfortable, avoided specific topics)	
therapist encouraged client to talk about whatever came to client's mind	
therapist taught client specific new skills or behaviors (relax muscles, control emotions, be assertive with others, act in social situations)	
therapist encouraged client to think about, view, or touch things that they are afraid of	
therapist reviewed or assigned homework exercises (writing down certain thoughts/feelings outside session, practice certain behaviours)	

therapist pointed out recurring themes or problems in client's relationships

therapist talked about the function or purpose that client's problem might have (e.g., avoid responsibility, keep others away from me)

therapist encouraged client to explore explanations for events or behaviors other than those that first came to client's mind

therapist made connections between the way client acts or feels toward therapist and the way that they act or feel in their other relationships

therapist encouraged client to see the choices they have in their life

therapist and client discussed client's dreams, fantasies, or wishes

therapist encouraged client to consider the positive and negative consequences of acting in a new way

therapist tried to help client identify the consequences (positive or negative) of client's behavior

therapist gave client advice or suggested practical solutions for client's problem

therapist shared personal information with client

therapist often explained what he/she was trying to do

therapist led the discussion most of the time

therapist focused on how disagreements between certain parts of client's personality have caused client's problems

therapist encouraged client to change specific behaviors

therapist focused on the ways client copes with client's problems

therapist encouraged client to look for evidence in support of or against one of client's beliefs or assumptions

therapist explored client's feelings about therapy

therapist encouraged client to view client's problem from a different perspective

therapist encouraged client to explore the personal meaning of an event or a feeling

therapist often focused on client's childhood experiences

therapist encouraged client to list the advantages and disadvantages of a belief or general rule that they follow

therapist had client role-play (act out or rehearse) certain scenes or situations

therapist tried to help client better understand how they relate to others, how this style of relating developed, and how it causes client's problems

therapist seemed interested in trying to understand what they were experiencing

therapist encouraged client to focus on client's moment-to-moment experience

therapist tried to help client better understand how their problem was due to certain beliefs or rules that they follow
therapist encouraged client to question their beliefs or to discover flaws in client's reasoning
therapist focused on a specific concern in client's relationships (disagreements/conflicts, major changes, loss of loved one, loneliness)
therapist encouraged client to explore ways in which they could make changes in their relationships (resolve a conflict in a relationship, fulfil a need, establish new relationships/contact old friends, avoid problems they had experienced in previous relationships)
therapist reviewed the gains client had made while in therapy
therapist reviewed the difficulties that client was currently experiencing
therapist encouraged client to examine their relationships with others (positive and negative aspects of their relationships, what client wants and what others want of client, the way client acts in relationships)
therapist encouraged client to think about ways in which they might prepare for major upcoming changes in their relationships (learning new skills, finding new friends)
therapist both accepted client for who they are and encouraged them to change
therapist encouraged client to identify situations in which client's feelings were invalidated (e.g., when a significant other told them their feelings were incorrect, situations in which they had strong feelings that seemed inappropriate)
therapist encouraged client to think about or be aware of things in client's life without judging them
therapist made it clear that client's problem was a treatable medical condition
therapist tried to help client better understand how their problems were due to difficulties in their social relationships

Appendix G: Researcher Allegiance Coding Criteria

Adapted from Yulish (2017). Modification in that dose is treated as separate moderator and so points range from -3 to +3 (removed extra +1 item about dose).

Points assigned as follows:

+1	If author advocates for treatment or developed treatment (Advocated for general family if between families; specific package if within family and if possible to ID)
+1	If #1 is true, and authors supervised the therapists, were the therapists in their own condition, or the therapists were extensively trained in the treatment
+1	If therapists received more supervision/training than other treatment
-1	If supervisor is not a recognized expert in treatment.
-1	If treatment protocol manual was altered by removing ingredient(s) or changing order in a theoretically deleterious manner.
-1	If therapists were prohibited from responding what reasonable therapist would routinely do AND prohibition is judged to be deleterious to treatment

Points summed to give total score from -3 to +3 for each treatment. Balance of allegiance—the moderator for this study—is determined by subtracting Allegiance A from Allegiance B.

"Advocates" for Therapy A = lead author(s) publishes disproportionately on Therapy A, publishes books on Therapy A, belongs to professional organization for Therapy A, publishes articles/op eds/opinion pieces in favour of Therapy A, gives interviews/public outreach in favour of Therapy A.

Appendix H: Characteristics of Primary Studies

Study	N	Age	Country	Race	Gender	Disorder	Within or Between Families	Comparison
Afshari, Hasani, 2020	68	27.60	Iran	0.00	41.00	Anxiety disorder	0	CBT-CBT
Agras, Schneider, Arnow, Raeburn, Telch, 1989	77	29.20	USA	x	0.00	Eating disorder	0	CBT-CBT
Agras, Walsh, Fairburn, Wilson, Kraemer, 2000	220	28.10	USA	77.00	0.00	Eating disorder	1	CBT-IPT
Arch, Eifert, Davies, Plumb Vilardaga, Rose, Craske, 2012	128	37.93	USA	67.00	48.00	Anxiety disorder	0	CBT-CBT
Arntz, Van Den Hout, 1996	36	34.10	Netherlands	x	61.00	Anxiety disorder	0	CBT-CBT
Barber, Mildrod, Gallop, Solomonov, Rudden, McCarthy, Chambless, 2020	116	x	USA	72.56	37.70	Anxiety disorder	1	CBT-PDT
Barlow, Craske, Cerny, Klosko, 1989	41	34.85	USA	x	29.27	Anxiety disorder	0	CBT-CBT
Barlow, Farchione, Bullis, Gallagher, Murray-Latin, Sauer-Zavala, Bentley, Thompson-Hollands, Conklin, Boswell, Ametaj, Carl, Boettcher, Cassiello-Robbins, 2017	179	30.70	USA	83.20	44.70	Anxiety disorder	0	CBT-CBT
Barlow, Rapee, Brown, 1992	65	x	USA	x	x	Anxiety disorder	1	CBT-EHT
Barrowclough, Colville, Russell, Burns, Tarrier, 2001	55	72.00	UK	x	23.00	Personality disorder	1	CBT-IPT
Bellino, Zizza, Rinaldi, Bogetto, 2007	26	30.55	Italy	x	26.92	Obsessive-compulsive/related disorder	0	CBT-CBT
Belloch, Cabello, Carrio, 2008	29	32.00	Spain	x	37.90	Depressive disorder	1	CBT-IPT
Bernecker, Constantino, Atkinson, Bagby, Ravitz, and McBride, 2016	69	38.67	Canada	82.81	24.64	Anxiety disorder	1	CBT-PDT
Beutel, Scheurich, Knebel, Michal, Wiltink, Graf-Morgenstern, Tschan, Milrod, Wellek, Subic-Wrana, 2013	54	36.22	Germany	x	42.60	Addictive disorder	1	CBT-EHT
Beutler, Moleiro, Malik, Harwood, Romanelli, Gallagher-Thompson, Thompson, 2003	40	33.06	USA	0.75	0.57	Anxiety disorder	0	CBT-CBT
Blakey, Abramowitz, Buchholz, Jessup, Jacoby, Reuman, Pentel, 2019	60	31.52	USA	0.68	0.15	Depressive disorder	1	CBT-IPT
Blanco, Markowitz, Hellerstein, Nezu, Wall, Olfson, Chen, Levenson, Onishi, Varona, Okuda, Hershman, 2019	134	52.70	USA	0.22	0.00	Depressive disorder	1	CBT-IPT
Bodenmann, Plancherel, Beach, Widmer, Gabriel, Meuwly, Charvoz, Hautzinger, Schramm, 2008	40	46.14	Switzerland	x	0.38	Anxiety disorder	0	CBT-CBT
Borkovec, Costello, 1993	66	37.50	USA	0.93	0.35	Anxiety disorder	0	CBT-EHT
Borkovec, Costello, 1993	66	37.50	USA	0.93	0.35	Anxiety disorder	1	CBT-CBT
Borkovec, Mathews, 1988	32	32.90	USA	x	0.43	Anxiety disorder	1	CBT-EHT
Borkovec, Mathews, 1988	32	32.90	USA	x	0.43	Anxiety disorder	0	CBT-CBT
Borkovec, Newman, Pincus, Lytle, 2002	76	37.14	USA	0.90	0.35	Sleep disorder	1	CBT-CBT
Bramoweth, Lederer, Youk, Germain, Chinman, 2020	63	55.10	USA	79.40	0.91			

Study	N	Age	Country	Race	Gender	Disorder	Within or Between Families	Comparison
Bruijniks, Lemmens, Hollon, Peeters, Cuijpers, Arntz, Dingemanse, Willems, van Oppen, Twisk, van den Boogaard, Spijker, Bosmans, Huibers, 2020	70	39.46	USA	0.83	0.46	Depressive disorder	1	CBT-IPT
Bruijniks, Lemmens, Hollon, Peeters, Cuijpers, Arntz, Dingemanse, Willems, van Oppen, Twisk, van den Boogaard, Spijker, Bosmans, Huibers, 2020	60	35.80	USA	0.88	0.38	Depressive disorder	1	CBT-IPT
Bryant, Mastrodomenico, Felmingham, Hopwood, Kenny, Kandris, Caholl, Creamer, 2008	58	35.19	Australia	x	0.48	Trauma disorder	0	CBT-CBT
Bryant, Mastrodomenico, Hopwood, Kenny, Cahill, Kandris, Taylor, 2013	45	33.85	Australia	x	x	Trauma disorder	0	CBT-CBT
Bryant, Moulds, Guthrie, Dang, Nixon, 2003	55	23.15	USA	63.64	56.36	Trauma disorder	0	CBT-CBT
Bryant, Sackville, Dang, Moulds, Guthrie, 1999	60	32.57	USA	100.00	0.83	Addictive disorder	0	CBT-CBT
Buckner, Zvolensky, Ecker, Schmidt, Lewis, Paulus, Lopez-Gamundi, Crapanzano, Bakhshaie, 2019	60	32.57	USA	100.00	0.83	Addictive disorder	0	CBT-CBT
Budney, Higgins, Radonovich, Novy, 2000	60	32.40	USA	93.50	0.75	Addictive disorder	0	CBT-Integrative
Budney, Higgins, Radonovich, Novy, 2000	39	40.05	UK	x	0.00	Addictive disorder	0	CBT-CBT
Budney, Moore, Rocha, Higgins, 2006	100	39.65	USA	44.00	95.00	Anxiety disorder	0	CBT-CBT
Burke, Drummond, Johnson, 1997	57	33.35	UK	x	0.14	Addictive disorder	1	CBT-CBT
Burling, Snyder Burling, Latini, 2001	141	35.94	Germany	x	0.34	Anxiety disorder	0	CBT-CBT
Butler, Fennell, Robson, Gelder, 1991	21	51.72	Italy	x	0.10	Trauma disorder	0	CBT-Integrative
Butollo, Karl, Konig, Rosner, 2016	165	35.60	New Zealand	x	0.27	Trauma disorder	0	CBT-EMDR
Capezzani, Ostacoli, Cavallo, Carletto, Fernandez, Solomon, Pagani, Cantelmi, 2013	100	38.35	New Zealand	0.84	0.31	Depressive disorder	0	CBT-IPT
Carter, Crowe, Carlyle, Frampton, Jordan, McIntosh, O'Toole, Whitehead, Joyce, 2012	62	31.95	UK	0.89	0.56	Depressive disorder	1	CBT-Integrative
Carter, McIntosh, Jordan , Porter, Frampton, Joyce, 2013	64	34.60	UK	x	0.22	Anxiety disorder	1	CBT-CBT
Clark, Ehlers, Hackmann, McManus, Fennell, Grey, Waddington, Wild, 2006	48	34.00	UK	x	0.33	Anxiety disorder	1	CBT-CBT
Clark, Salkovskis, Hackmann, Middleton, Anastasiades, Gelder, 1994	237	36.20	USA	0.51	0.25	Somatic disorder	1	CBT-CBT
Clark, Salkovskis, Hackmann, Wells, Femmell, Ludgate, Ahmad, Richards, Gelder, 1998	22	47.45	USA	0.64	0.32	Depressive disorder	0	CBT-PDT
Connolly Gibbons, Gallop, Thompson, Luther, Crits-Christoph, Jacobs, Yin, Crits-Christoph, 2006	42	41.04	France	x	0.24	Depressive disorder	0	CBT-CBT
Constantino, Marnell, Haile, Kanther-Sista, Wolman, Zappert, Arnow, 2008	65	35.77	France	x	0.26			

Study	N	Age	Country	Race	Gender	Disorder	Within or Between Families	Comparison
Cottraux, Note ,Yao, Lafont, Note, Mollarda, Bouvard, Sauteraud, Bourgeois, Dartigues, 2001	96	35.00	USA	52.70	0.33	Obsessive-compulsive/related disorder	1	CBT-CBT
Cottraux, Note, Yao, de Mey-Guillard, Bonasse, Djamoussian, Mollard, Note, Chen, 2008	243	33.20	USA	57.90	0.77	Trauma disorder	0	CBT-EHT
Craske, Meuret, Ritz, Treanor, Dour, Rosenfield, 2019	70	25.90	Copenhagen	1.00	0.01	Anxiety disorder	0	CBT-CBT
Crits-Christoph, Siqueland, Blaine, Frank, Luborsky, Onken, Muenz, Thase, Weiss, Gastfriend, Woody, Barber, Butler, Daley, Salloum, Bishop, Najavits, Lis, Mercer, Griffin, Moras, Beck, 1999	30	35.10	USA	x	0.54	Addictive disorder	1	CBT-PDT
Daniel, Poulsen, Lunn, 2016	40	34.00	Netherlands	x	0.40	Eating disorder	0	CBT-PDT
de Ruiter, Rijken, Garssen, Kraaimaat, 1989	241	39.90	USA	81.70	0.44	Anxiety disorder	1	CBT-CBT
Deckersbach, Rauch, Buhlmann, Wilhelm, 2006	341	38.91	Netherlands	55.00	0.30	Neurodevelopmental disorder Depressive disorder	0 1	CBT-Integrative CBT-CBT
Dimidjian, Hollon, Dobson, Schmaling, Kohlenberg, Addis, Gallop, McGlinchey, Markley, Gollan, Atkins, Dunner, Jacobson, 2006	65	38.50	Canada	91.00	0.33	Depressive disorder	1	CBT-PDT
Driessen, Van, Don, Peen, Kool, Westra, Hendriksen, Schoevers, Cuijpers, Twisk, Dekker, 2013	110	39.00	UK	x	0.32	Anxiety disorder	0	CBT-CBT
Dugas, Brillon, Savard, Turcotte, Gaudet, Ladouceur, Leblanc, Gervais, 2010	110	39.00	UK	x	0.32	Anxiety disorder	0	CBT-CBT
Durham, Murphy, Allan, Richard, Treliving, Fenton, 1994	91	39.67	UK	0.70	0.42	Anxiety disorder	0	CBT-PDT
Durham, Murphy, Allan, Richard, Treliving, Fenton, 1994	43	38.19	Canada	0.77	0.42	Trauma disorder	1	CBT-Integrative
Ehlers, Hackman, Grey, Wild, Liness, Albert, Deale, Stott, Clark, 2014	21	x	Netherlands	x	0.14	Depressive disorder	0	EHT-EHT
Ellison, Greenberg, Goldman, Angus, 2009	44	34.30	Netherlands	x	0.47	Obsessive-compulsive/related disorder	1	CBT-CBT
Emmeikamp, Visser, Hoekstra, 1988	18	29.90	Netherlands	x	0.50	Obsessive-compulsive/related disorder	0	CBT-CBT
Emmelkamp, Beens, 1991	33	43.97	Netherlands	x	0.55	Personality disorder	1	CBT-Integrative
Emmelkamp, Benner, Kuipers, Feiertag, Koster, van Apeldoorn, 2006	130	25.90	UK	0.95	0.98	Anxiety disorder	0	CBT-CBT
Emmelkamp, Krijn, Hulsbosch, de Vries, Schuemie, van der Mast, 2002	60	49.39	Iran	x	0.32	Eating disorder	0	CBT-IPT
Fairburn, Bailey-Straebler, Basden, Doll, Jones, Murphy, O'Connor, Cooper, 2015	57	43.90	Switzerland	x	0.14	Trauma disorder	1	CBT-CBT
Farahimanesh, Moradi, Sadeghi, Jobson, 2021	96	34.90	USA	0.63	0.00	Anxiety disorder	0	CBT-CBT
Flückiger, Forrer, Schnider, Bättig,	45	31.80	USA	0.73	0.00			

Study	N	Age	Country	Race	Gender	Disorder	Within or Between Families	Comparison
Bodenmann, Zinbarg, 2016						Trauma disorder	0	CBT-CBT
Foa, Dancu, Hembree, Jaycox, Meadows, Street, 1999	179	31.30	USA	0.49	0.00	Trauma disorder	0	CBT-CBT
Foa, Hembree, Cahill, Rauch, Riggs, Feeny, Yadin, 2005	93	38.25	UK	0.74	0.32	Trauma disorder	1	CBT-CBT
Foa, Rothbaum, Riggs, Murdock, 1991	92	24.20	UK	x	0.00	Depressive disorder	0	CBT-PDT
Fonagy, Lemma, Target, O'Keeffe, Constantinou, Ventura Wurman, Luyten, Allison, Roth, Cape, Pilling, 2020	50	27.20	Sweden	x	x	Eating disorder	0	CBT-CBT
Freeman, Barry, Dunkeld-Turnbull, Henderson, 1988	86	30.60	Netherlands	x	0.93	Eating disorder	1	CBT-CBT
Ghaderi, 2006	369	35.52	Germany	x	0.24	Personality disorder	0	Integrative-PDT
Giesen-Bloo, Van Dyck, Spinhoven, Van Tilburg, Dirksen, Van Asselt,Kremers, Nadort, Arntz, 2006	34	39.64	Canada	x	0.26	Anxiety disorder	0	CBT-CBT
Gloster, Hauke, Höfler. Einsle, Fydrich, Hamm, Sthröhle, Wittchen, 2013	147	40.46	Switzerland	x	0.44	Depressive disorder	1	EHT-EHT
Greenberg, Watson, 1998	114	40.25	USA	0.97	0.47	Depressive disorder	0	CBT-Integrative
grosse Holtforth, Krieger, Zimmermann, Altenstein-Yamanaka, Dorig, Meisch, Hayes, 2019	81	32.92	USA	80.20	0.35	Depressive disorder	0	CBT-PDT
Hardy, Barkham, Shapiro, Stiles, Rees, Reynolds, 1995	30	29.50	Sweden	x	0.37	Anxiety disorder	1	CBT-CBT
Hayes-Skelton, Roemer, Orsillo, 2013	50	40.22	Brazil	0.28	0.10	Anxiety disorder	1	CBT-CBT
Hellstrom, Fellenius, Ost, 1996	125	30.40	USA	0.90	0.00	Depressive disorder	0	CBT-CBT
Hemanny, Carvalho ,Maia, Reis, Botelho, Bonavides, Seixas, de Oliveira, 2020	75	36.21	USA	36.00	0.00	Depressive disorder	0	Integrative-IPT
Hersen, Bellack, Himmelhoch, Thase, 1984	82	43.67	USA	0.15	0.67	Addictive disorder	0	CBT-CBT
Hien, Cohen, Miele, Litt, Capstick, 2004	62	45.20	Norway	x	0.75	Addictive disorder	1	CBT-CBT
Hien, Smith, Owens, Lopez-Castro, Ruglass, Papini, 2018	80	55.40	USA	0.93	0.00	Trauma disorder	0	CBT-CBT
Hoffart, Oktedalen, Langkaas, Wampold, 2013	84	39.00	Netherlands	x	0.35	Depressive disorder	0	CBT-CBT
Hopko, Armento, Robertson, Ryba, Carvalho, Colman, Mullane, Gawrysiak, Bell, McNulty, Lejuez, 2011	68	45.40	Germany	x	0.29	Anxiety disorder	0	CBT-EMDR
Horst, Den Oudsten, Zijlstra, de Jongh, Lobbestael, De Vries, 2017	100	33.00	Germany	x	0.29	Anxiety disorder	0	CBT-CBT
Hoyer, Beesdo, Gloster, Runge, Höfler, Becker, 2009	80	37.00	Brazil	0.83	0.36	Depressive disorder	1	PDT-PDT
Huber, Henrich, Gastner, Klug, 2012	60	46.05	Italy	1.00	0.00	Anxiety disorder	0	CBT-CBT
Ito, de Araujo, Tess, de Barros-Nero, Asbahr, Marks, 2001	151	37.96	USA	0.85	0.25	Eating disorder	0	CBT-CBT
Jackson, Pietrabissa, Rossi, Manzoni, Castelnuovo, 2018	40	38.58	Netherlands	x	0.30			

Study	N	Age	Country	Race	Gender	Disorder	Within or Between Families	Comparison
Jacobson, Dobson, Truax, Addis, Koerner, Gollan, Gortner, Prince, 1996	48	31.60	Netherlands	x	0.06	Depressive disorder	0	CBT-CBT
Kampmann, Emmelkamp, Hartanto, Brinkman, Zijlstra, Morina, 2016	30	79.41	USA	0.73	0.30	Anxiety disorder	0	CBT-CBT
Keijsers, Maas, van Opdorp, van Minnen, 2016	395	45.94	USA	0.87	0.41	Obsessive-compulsive/related disorder	0	CBT-CBT
Kiosses, Arean, Teri, Alexopoulos, 2010	31	35.50	Canada	81.00	0.00	Depressive disorder	0	CBT-EHT
Klein, Leon, Li, D'Zurilla, Black, Vivian, Dowling, Arnow, Manber, Markowitz, Kocsis, 2011	22	43.45	Canada	22.00	0.41	Depressive disorder	0	CBT-EHT
Koszycki, Bisserbe, Blier, Bradwejn Markowitz, 2012	74	32.75	Switzerland	x	0.31	Depressive disorder	1	EHT-IPT
Koszycki, Raab, Aldosary, Bradwejn, 2010	61	34.21	Netherlands	x	0.16	Anxiety disorder	1	CBT-CBT
Kramer, Kolly, Berthoud, Keller, Preisig, Caspar, Berger, de Roten, Marquet, Despland, 2014	65	45.20	Norway	x	0.42	Personality disorder	1	PDT-PDT
Kunze, Arntz, Morina, Kindt, Lancee, 2017	416	34.61	Germany	x	0.46	Sleep disorder	0	CBT-CBT
Langkaas, Hoffart, Oktedalen, Ulvenes, Hembree, Smucker, 2017	57	42.50	Germany	x	0.19	Trauma disorder	0	CBT-CBT
Leichsenring, Salzer, Beutel, Herpertz, Hiller, Hoyer, Huesing, Joraschky, Nolting, Poehlmann, Ritter, Stangier, Strauss, Stuhldreher, Tefikow, Teismann, Willutzki, Wiltink, Leibing, 2013	70	34.69	USA	0.50	0.57	Anxiety disorder	0	CBT-PDT
Leichsenring, Salzer, Jaeger, Kachele, Kreische, Leweke, Ruger, Winkelbach, Leibing, 2009	26	38.40	USA	0.69	0.69	Anxiety disorder	0	CBT-Integrative
Lemmens, Arntz, Peeters, Hollon, Roefs, Huibers, 2015	110	40.10	USA	0.65	0.30	Depressive disorder	0	CBT-IPT
Lipsitz, Gur, Vermes, Petkova, Cheng, Miller, Laino, Leibowitz, Fyer, 2008	110	40.10	USA	0.65	0.30	Anxiety disorder	1	EHT-IPT
Markowitz, Kocsis, Christos, Bleiberg, Carlin, 2008	87	38.00	UK	x	0.64	Depressive disorder	1	EHT-IPT
Markowitz, Petkova, Neria, Van Meter, Zhao, Hembree, Lovell, Biyanova, Marshall, 2015	74	39.71	USA	0.47	0.00	Trauma disorder	1	CBT-CBT
Markowitz, Petkova, Neria, Van Meter, Zhao, Hembree, Lovell, Biyanova, Marshall, 2015	36	19.30	USA	0.83	0.36	Trauma disorder	1	CBT-IPT
Marks, Lovell, Noshirvani, Livanou, Thrasher, 1998	49	40.41	USA	0.04	0.57	Trauma disorder	1	CBT-CBT
McDonagh, Friedman, McHugo, Ford, Sengupta, Mueser, Demment, Fournier, Schnurr, Descamps, 2005	34	35.60	Netherlands	x	0.68	Trauma disorder	0	CBT-CBT
McIndoo, File, Preddy, Clark, Hopko, 2016	89	36.30	USA	44.90	1.00	Depressive disorder	1	CBT-CBT
McKay, Lynch, Coviello, Morrison, Cary, Skalina, Plebani, 2010	128	41.33	USA	0.90	0.47	Addictive disorder	0	CBT-CBT

Study	N	Age	Country	Race	Gender	Disorder	Within or Between Families	Comparison
Mersch, 1995	24	39.40	USA	0.97	0.30	Anxiety disorder	0	CBT-CBT
Morgenstern, Irwin, Wainberg, Parsons, Muench, Bux, Kahler, Marcus, Schulz-Heik, 2007	140	37.80	Netherlands	0.65	0.44	Addictive disorder	0	Integrative-Integrative
Muran, Safran, Wallner Samstag, Winston, 2005	30	40.63	Australia	x	0.53	Personality disorder	0	CBT-PDT
Murphy, Carney, Knesevich, Wetzel, Whitworth, 1995	60	37.73	Norway	0.97	0.28	Depressive disorder	0	CBT-CBT
Nijdam, Gersons, Reitsma, de Jongh, Olff, 2012	65	45.18	Norway	x	0.43	Trauma disorder	1	CBT-EMDR
Nixon, 2012	20	30.00	Sweden	x	0.30	Trauma disorder	0	CBT-Integrative
Nordahl, Borkovec, Hagen, Kennair, Hjemdal, Solem, Hansen, Haseth, Wells, 2018	36	32.60	Sweden	x	0.33	Anxiety disorder	1	CBT-CBT
Øktedalen, Hoffart, Langkaas, 2015	18	37.50	Sweden	x	0.22	Trauma disorder	1	CBT-CBT
Ost & Westling, 1995	46	41.30	Sweden	x	0.09	Anxiety disorder	0	CBT-CBT
Ost, 1988	16	37.90	Sweden	x	0.81	Anxiety disorder	0	CBT-CBT
Ost, Alm, Brandberg, Breitholtz, 2001	48	40.61	UK	x	0.58	Anxiety disorder	0	CBT-CBT
Ost, Fellenius, Sterner, 1991	108	30.21	USA	0.64	0.95	Anxiety disorder	0	CBT-CBT
Paunovic, Ost, 2000	171	31.99	USA	0.71	0.00	Trauma disorder	0	CBT-CBT
Power, McGoldrick, Brown, Buchanan, Sharp, Swanson, Karatzias, 2002	440	43.50	UK	0.91	0.34	Trauma disorder	0	CBT-EMDR
Reger, Koenen-Woods, Zetocha, Smolenski, Holloway, Rothbaum, Difede, Rizzo, Edwards-Stewart, Skopp, Mishkind, Reger, Gahm, 2016	62	41.18	Australia	x	47.00	Trauma disorder	0	CBT-CBT
Resick, Nishith, Waver, Asin, Feuer, 2002	30	30.50	Netherlands	x	0.47	Trauma disorder	1	CBT-CBT
Richards,Ekers, McMillan, Taylor, Byford, Warren, Barrett, Farrand, Gilbody, Kuyken, O'Mahen, Watkins, Wright, Hollon, Reed, Rhodes, Fletcher, Finning, 2016	268	44.94	Germany	x	0.34	Depressive disorder	0	CBT-CBT
Sannibale, Teesson, Creamer, Sitharthan, Bryant, Sutherland, Taylor, Bostock-Matusko,Visser, Peek-O'Leary, 2013	103	39.88	Israel	x	0.44	Addictive disorder	0	CBT-CBT
Scholing, Emmelkamp, 1993	43	34.70	USA	x	0.18	Anxiety disorder	0	CBT-CBT
Schramm, Kriston, Zobel, Bailer, Wambach, Backenstrass, Klein, Schoepf, Schnell, Gumz, Bausch, Fangmeier, Meister, Berger, Hautzinger, Harter, 2017	108	39.20	USA	0.87	0.57	Depressive disorder	0	EHT-Integrative
Shalev, Ankri, Israeli-Shalev, Peleg, Adessky, Freedman, 2012	30	39.90	USA	0.47	0.53	Trauma disorder	1	CBT-CBT
Shear, Pilkonis, Cloitre, Leon, 1994	126	43.90	USA	0.55	0.52	Anxiety disorder	0	CBT-EHT
Simpson, Foa, Liebowitz, Ledley, Huppert, Cahill, Vermes, Schmidt, Hembree, Franklin,	161	39.00	USA	0.74	0.35	Obsessive-compulsive/related disorder	1	CBT-CBT

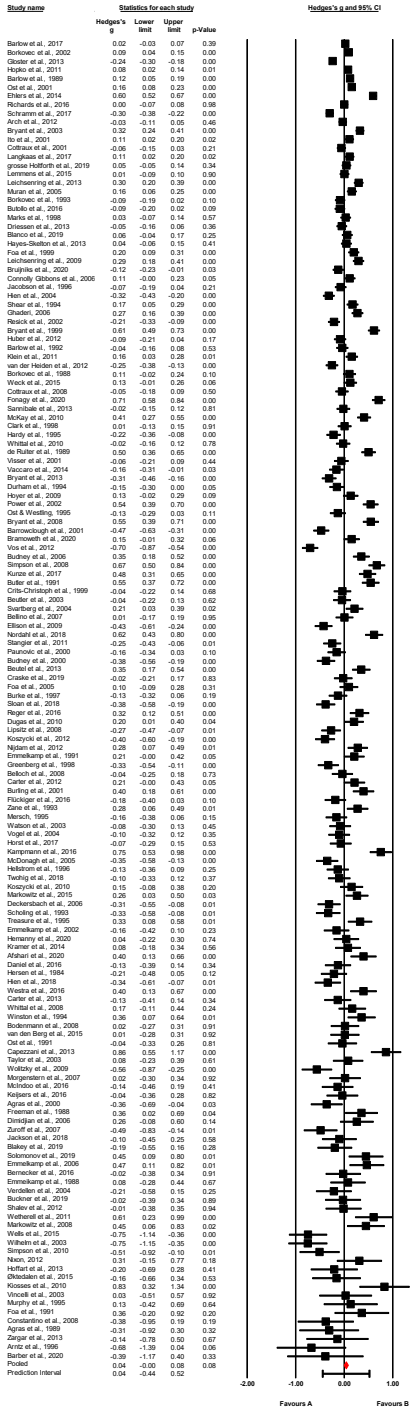
Study	N	Age	Country	Race	Gender	Disorder	Within or Between Families	Comparison
Campeas, Hahn, Petkova, 2008						Obsessive-compulsive/related disorder	0	CBT-CBT
Simpson, Zuckoff, Maher, Page, Franklin, Foa, Schmidt, Wang, 2010	76	34.25	Germany	x	0.49	Trauma disorder	0	CBT-CBT
Sloan, Marx. Lee, Resick, 2018	50	34.00	Norway	100.00	50.00	Anxiety disorder	0	CBT-PDT
Solomonov, Falkenstrom, Gorman, McCarthy, Milrod, Rudden, Chambless, Barber, 2019	60	37.00	Canada	77.00	25.00	Anxiety disorder	1	CBT-IPT
Stangier, Schramm,Heidenreich, Berger, Clark, 2011	60	37.00	Canada	77.00	25.00	Personality disorder	1	CBT-PDT
Svartberg, Stiles, Seltzer, 2004	30	25.02	UK	x	x	Trauma disorder	1	CBT-CBT
Taylor, Thordarson, Maxfield, Fedoroff, Lovell, Ogrodniczuk, 2003	58	27.25	USA	0.78	0.31	Trauma disorder	1	CBT-EMDR
Taylor, Thordarson, Maxfield, Fedoroff, Lovell, Ogrodniczuk, 2003	50	38.24	Australia	x	0.56	Eating disorder	0	CBT-Integrative
Treasure, Todd, Brolly, Tiller, Nehmed, Denman, 1995	151	41.25	Netherlands	x	0.66	Obsessive-compulsive/related disorder	1	CBT-CBT
Twohig, Abramowitz, Smith, Fabricant, Jacoby, Morrison, Bluett, Reuman, Blakey, Ledermann, 2018	104	36.28	Netherlands	x	0.37	Obsessive-compulsive/related disorder	0	CBT-CBT
Vaccaro, Jones, Menzies, Wootton, 2014	96	39.55	Netherlands	x	0.41	Trauma disorder	1	CBT-EMDR
van den Berg, de Bont, van der Vleugel, de Roos, de Jongh, Van Minnen, van der Gaag, 2015	108	41.48	Netherlands	0.73	0.44	Anxiety disorder	1	CBT-CBT
van der Heiden,Muris, van der Molen, 2012	126	35.00	Netherlands	x	0.27	Neurodevelopmental disorder	0	CBT-CBT
Verdellen, Keijsers, Cath, Hoogduin , 2004	43	20.57	Netherlands	x	0.79	Anxiety disorder	0	CBT-CBT
Vincelli, Anolli, Bouchard, Wiederhold, Zurloni, Riva, 2003	12	43.83	Italy	x	0.00	Somatic disorder	0	CBT-CBT
Visser, Bouman, 2001	78	36.20	Netherlands	x	0.50	Obsessive-compulsive/related disorder	0	CBT-CBT
Vogel, Stules, Gotestam, 2004	35	35.70	Norway	x	0.29	Anxiety disorder	0	CBT-IPT
Vos, Huibers, Diels, Arntz, 2012	91	34.79	Netherlands	x	0.23	Depressive disorder	1	CBT-EHT
Watson, Gordon, Stermac,Kalogerakos, Steckley, 2003	93	40.30	Canada	0.90	0.33	Somatic disorder	1	CBT-CBT
Weck, Neng, Richtberg, Jakob, Stangier, 2015	84	40.05	Germany	x	40.48	Trauma disorder	1	CBT-CBT
Wells, Walton, Lovell, Proctor, 2015	32	41.20	UK	x	0.63	Anxiety disorder	0	CBT-CBT
Westra, Constantino, Antony, 2016	85	33.33	Canada	0.75	0.12	Anxiety disorder	0	CBT-CBT
Wetherell, Afari, Ayers, Stoddard, Ruberg, Sorrell, Liu, Petkus, Thorp, Kraft, Patterson, 2011	16	70.80	USA	0.62	0.53	Obsessive-compulsive/related disorder	0	CBT-CBT
Whittal, Robichaud, Thordarson, McLean, 2008	41	36.10	Canada	0.85	0.37	Obsessive-compulsive/related disorder	0	CBT-CBT
Whittal, Woody, McLean, Rachman, Robichaud, 2010	73	31.50	Canada	0.85	0.53	Neurodevelopmental disorder	0	CBT-EHT
Wilhelm, Deckersbach, Coffey, Bohne, Peterson, Baer, 2003	29	34.91	USA	x	0.55			

Study	N	Age	Country	Race	Gender	Disorder	Within or Between Families	Comparison
Winston, Laikin, Pollack, Samstag, McCullough, Muran, 1994	81	40.80	USA	x	41.00	Personality disorder	0	PDT-PDT
Wolitzky, Telch, 2009	88	20.08	USA	0.49	0.31	Anxiety disorder	0	CBT-CBT
Zane, Williams, 1993	45	x	USA	x	0.31	Anxiety disorder	1	CBT-CBT
Zargar, Farid, Atef-Vahid, Afshar, Omidi, 2013	18	38.60	Iran	x	0.00	Anxiety disorder	0	CBT-CBT
Zuroff, Koestner, Moskowitz, McBride, Marshall, Bagby, 2007	95	42.01	Canada	x	0.31	Depressive disorder	0	CBT-IPT

Note. “Age” = mean age. “Race” = percentage of participants who are White, x = not reported. “Gender” = percentage of participants who are men, x = not reported. “Within or Between Families,” 0 = within the same therapy family, 1 = between therapy families. “Comparison,” CBT = cognitive behavioural therapies, EHT = experiential and humanistic therapies, EMDR = eye movement desensitization and reprocessing, IPT = interpersonal therapy, PDT = psychodynamic therapies

Appendix I: Forest Plot of Effect Sizes with Randomly Distributed Signs

Effect Sizes of Therapy A vs. Therapy B with Randomly Distributed Signs, Collapsed Across Timepoints and Outcomes

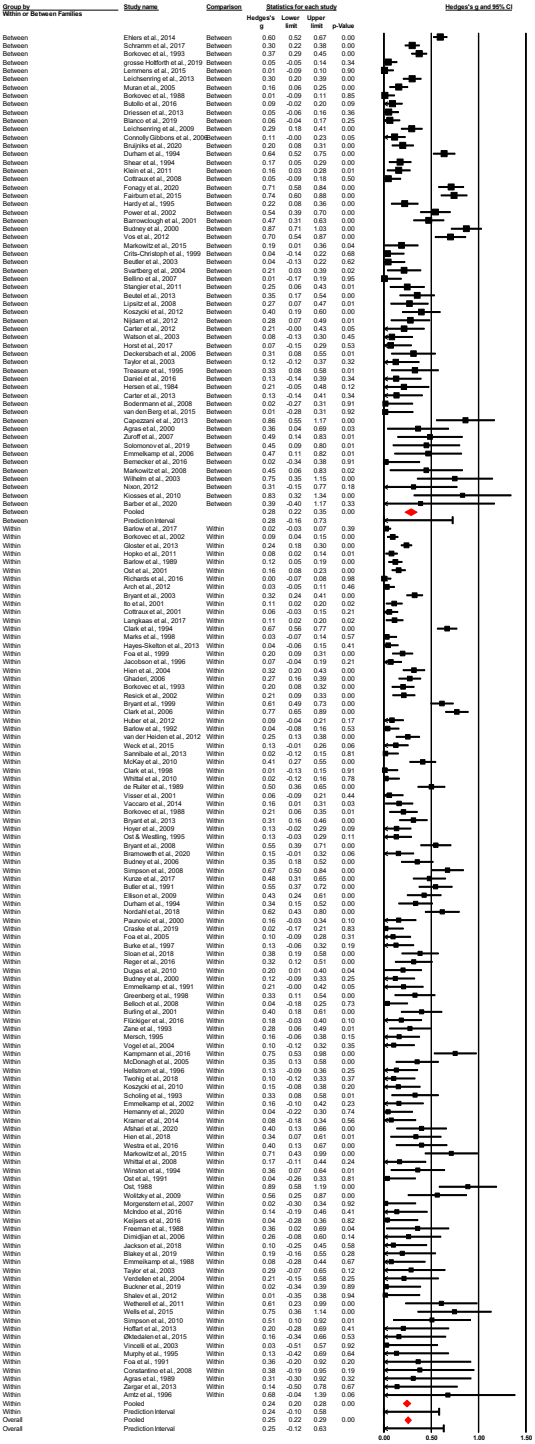


Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are sorted in descending order of precision (i.e., ascending order of variance). The red lozenge represents the summary effect size, with its width representing its 95% confidence range. All effect sizes were assigned positive or negative signs randomly; the summary effect is therefore expected to approximate zero. The prediction interval indicates the expected range of the effect size in 95% of comparable studies.

Appendix J: Forest Plots of Absolute Valued Effect Sizes

Figure J1

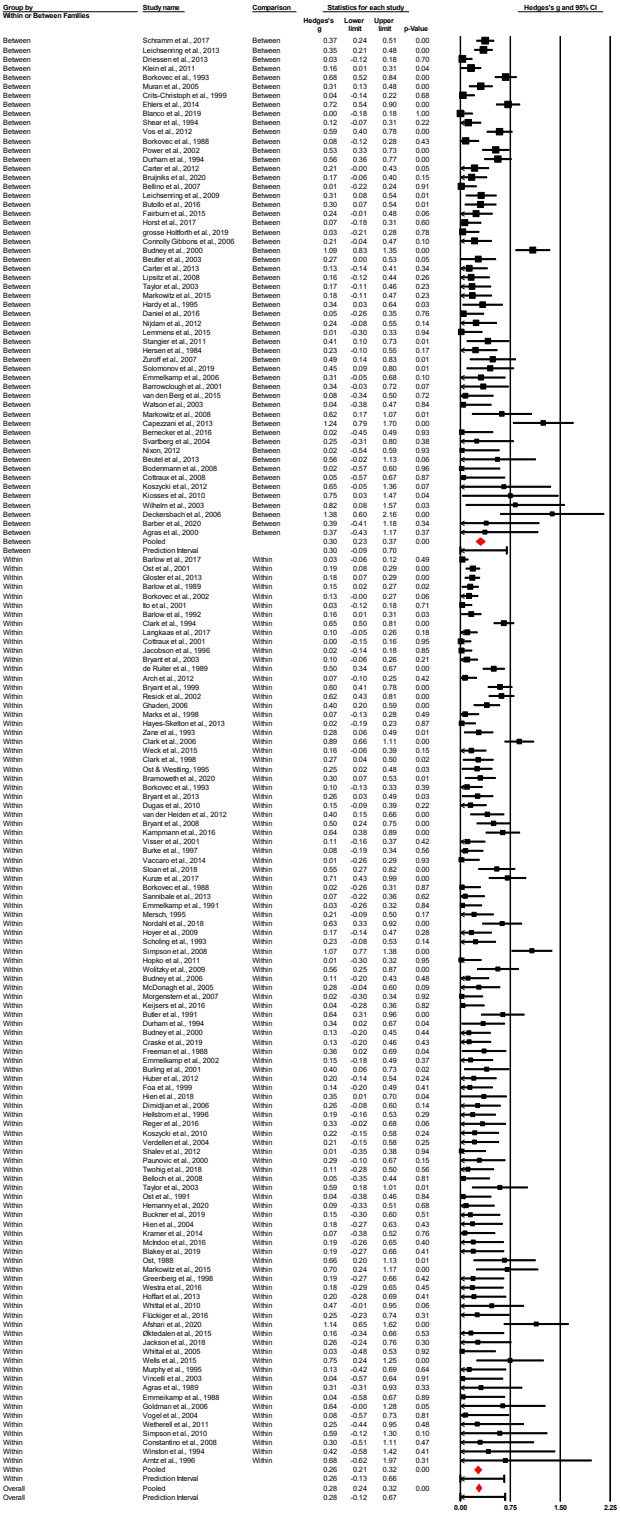
Forest Plot of Effect Sizes of Therapy A vs. Therapy B, Collapsed Across Timepoints and Outcomes



Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are grouped by comparison (within the same therapy family vs. between different therapy families) and sorted in descending order of precision (i.e., ascending order of variance). The red lozenges represents the summary effect sizes for within-family comparisons, between-family comparisons, and the overall summary effect; their widths represents their 95% confidence ranges. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. All effect sizes are absolute values (i.e., given positive signs); the forest plot therefore begins with zero as the minimum value and does not depict negative values for the confidence or prediction intervals. Outlying effects were not included.

Figure J2

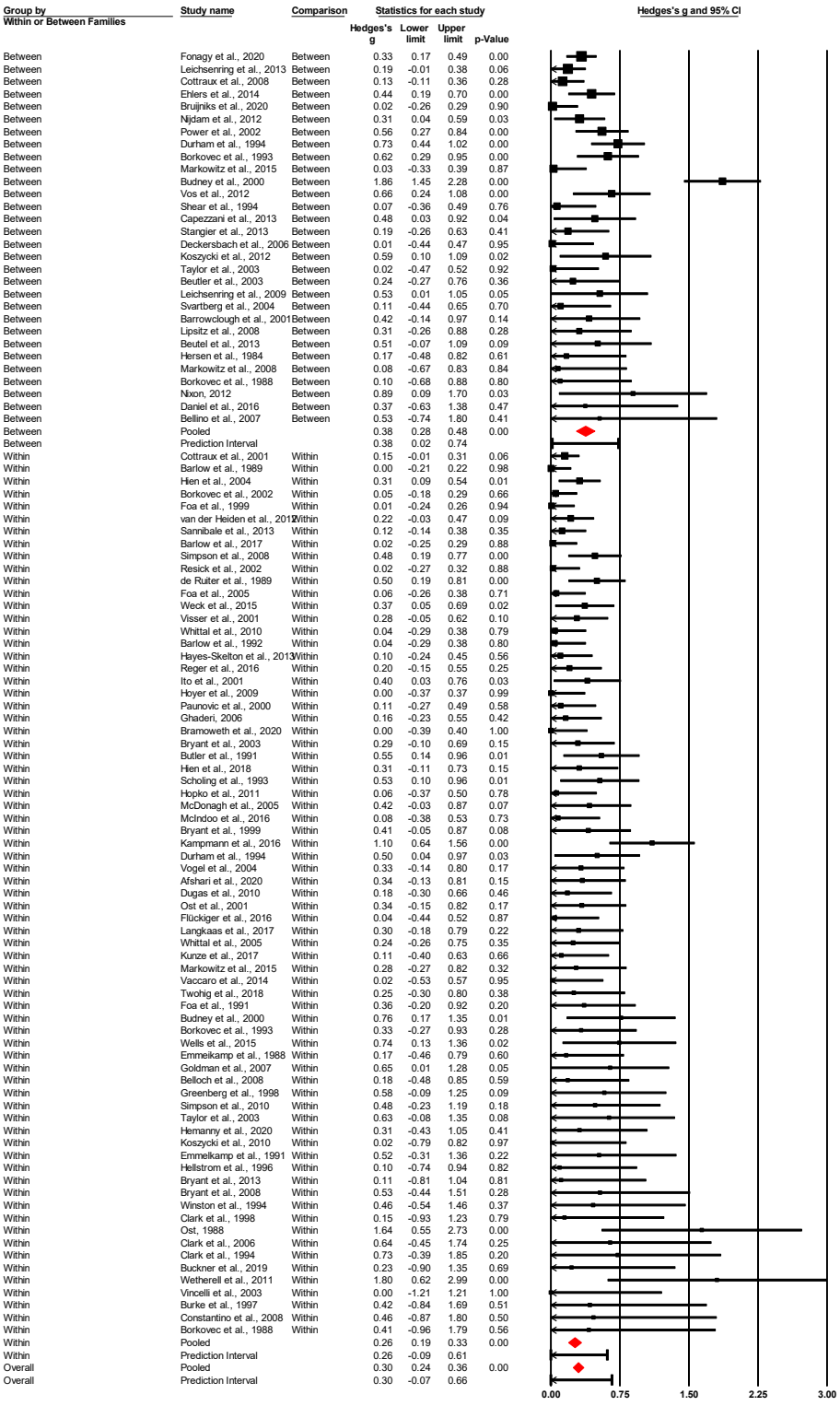
Forest Plot of Absolute Valued Effect Sizes of Therapy A vs. Therapy B, Primary Measures at Termination



Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are grouped by comparison (within the same therapy family vs. between different therapy families) and sorted in descending order of precision (i.e., ascending order of variance). The red lozenges represents the summary effect sizes for within-family comparisons, between-family comparisons, and the overall summary effect; their widths represents their 95% confidence ranges. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. All effect sizes are absolute values (i.e., given positive signs); the forest plot therefore begins with zero as the minimum value and does not depict negative values for the confidence or prediction intervals. Outlying effects were not included.

Figure J3

Forest Plot of Absolute Valued Effect Sizes of Therapy A vs. Therapy B, Secondary Measures at Termination



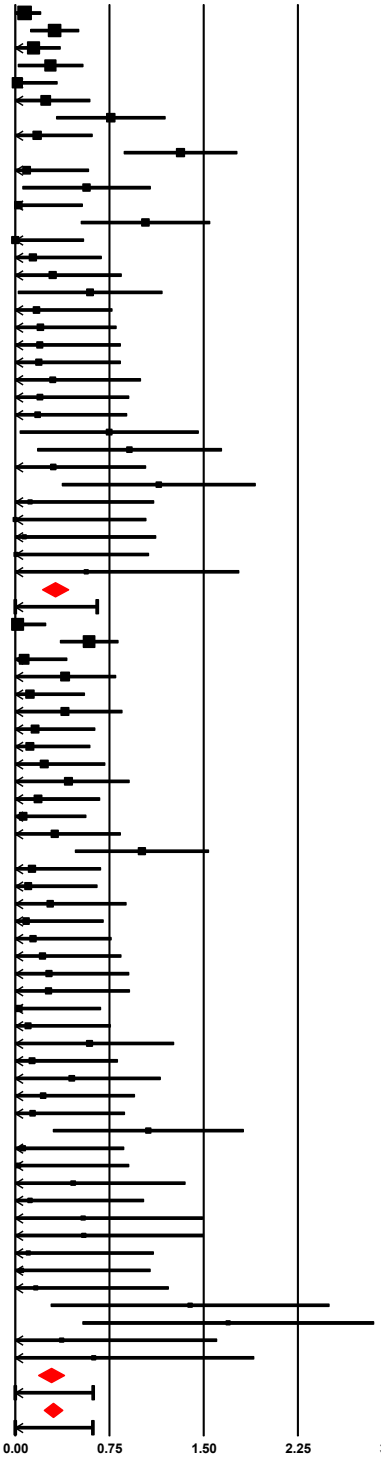
Group by Within or Between Families	Study name	Comparison	Hedges's g	Lower limit	Upper limit	p-Value
Between	Fonagy et al., 2020	Between	0.33	0.17	0.49	0.00
Between	Leichsenring et al., 2013	Between	0.19	-0.01	0.38	0.06
Between	Cottraux et al., 2008	Between	0.13	-0.11	0.36	0.28
Between	Ehlers et al., 2014	Between	0.44	0.19	0.70	0.00
Between	Bruijniks et al., 2020	Between	0.02	-0.26	0.29	0.90
Between	Nijdam et al., 2012	Between	0.31	0.04	0.59	0.03
Between	Power et al., 2002	Between	0.56	0.27	0.84	0.00
Between	Durham et al., 1994	Between	0.73	0.44	1.02	0.00
Between	Borkovec et al., 1993	Between	0.62	0.29	0.95	0.00
Between	Markowitz et al., 2015	Between	0.03	-0.33	0.39	0.87
Between	Budney et al., 2000	Between	1.86	1.45	2.28	0.00
Between	Vos et al., 2012	Between	0.66	0.24	1.08	0.00
Between	Shear et al., 1994	Between	0.07	-0.36	0.49	0.76
Between	Capezzani et al., 2013	Between	0.48	0.03	0.92	0.04
Between	Stangier et al., 2013	Between	0.19	-0.26	0.63	0.41
Between	Deckersbach et al., 2006	Between	0.01	-0.44	0.47	0.95
Between	Koszycki et al., 2012	Between	0.59	0.10	1.09	0.02
Between	Taylor et al., 2003	Between	0.02	-0.47	0.52	0.92
Between	Beutler et al., 2003	Between	0.24	-0.27	0.76	0.36
Between	Leichsenring et al., 2009	Between	0.53	0.01	1.05	0.05
Between	Svartberg et al., 2004	Between	0.11	-0.44	0.65	0.70
Between	Barrowclough et al., 2001	Between	0.42	-0.14	0.97	0.14
Between	Lipsitz et al., 2008	Between	0.31	-0.26	0.88	0.28
Between	Beutel et al., 2013	Between	0.51	-0.07	1.09	0.09
Between	Hersen et al., 1984	Between	0.17	-0.48	0.82	0.61
Between	Markowitz et al., 2008	Between	0.08	-0.67	0.83	0.84
Between	Borkovec et al., 1988	Between	0.10	-0.68	0.88	0.80
Between	Nixon, 2012	Between	0.89	0.09	1.70	0.03
Between	Daniel et al., 2016	Between	0.37	-0.63	1.38	0.47
Between	Bellino et al., 2007	Between	0.53	-0.74	1.80	0.41
Between	Pooled		0.38	0.28	0.48	0.00
Between	Prediction Interval		0.38	0.02	0.74	
Within	Cottraux et al., 2001	Within	0.15	-0.01	0.31	0.06
Within	Barlow et al., 1989	Within	0.00	-0.21	0.22	0.98
Within	Hien et al., 2004	Within	0.31	0.09	0.54	0.01
Within	Borkovec et al., 2002	Within	0.05	-0.18	0.29	0.66
Within	Foa et al., 1999	Within	0.01	-0.24	0.26	0.94
Within	van der Heiden et al., 2012	Within	0.22	-0.03	0.47	0.09
Within	Sannibale et al., 2013	Within	0.12	-0.14	0.38	0.35
Within	Barlow et al., 2017	Within	0.02	-0.25	0.29	0.88
Within	Simpson et al., 2008	Within	0.48	0.19	0.77	0.00
Within	Resick et al., 2002	Within	0.02	-0.27	0.32	0.88
Within	de Ruiter et al., 1989	Within	0.50	0.19	0.81	0.00
Within	Foa et al., 2005	Within	0.06	-0.26	0.38	0.71
Within	Weck et al., 2015	Within	0.37	0.05	0.69	0.02
Within	Visser et al., 2001	Within	0.28	-0.05	0.62	0.10
Within	Whittal et al., 2010	Within	0.04	-0.29	0.38	0.79
Within	Barlow et al., 1992	Within	0.04	-0.29	0.38	0.80
Within	Hayes-Skelton et al., 2013	Within	0.10	-0.24	0.45	0.56
Within	Reger et al., 2016	Within	0.20	-0.15	0.55	0.25
Within	Ito et al., 2001	Within	0.40	0.03	0.76	0.03
Within	Hoyer et al., 2009	Within	0.00	-0.37	0.37	0.99
Within	Paunovic et al., 2000	Within	0.11	-0.27	0.49	0.58
Within	Ghaderi, 2006	Within	0.16	-0.23	0.55	0.42
Within	Bramoweth et al., 2020	Within	0.00	-0.39	0.40	1.00
Within	Bryant et al., 2003	Within	0.29	-0.10	0.69	0.15
Within	Butler et al., 1991	Within	0.55	0.14	0.96	0.01
Within	Hien et al., 2018	Within	0.31	-0.11	0.73	0.15
Within	Scholing et al., 1993	Within	0.53	0.10	0.96	0.01
Within	Hopko et al., 2011	Within	0.06	-0.37	0.50	0.78
Within	McDonagh et al., 2005	Within	0.42	-0.03	0.87	0.07
Within	McIndoo et al., 2016	Within	0.08	-0.38	0.53	0.73
Within	Bryant et al., 1999	Within	0.41	-0.05	0.87	0.08
Within	Kampmann et al., 2016	Within	1.10	0.64	1.56	0.00
Within	Durham et al., 1994	Within	0.50	0.04	0.97	0.03
Within	Vogel et al., 2004	Within	0.33	-0.14	0.80	0.17
Within	Afshari et al., 2020	Within	0.34	-0.13	0.81	0.15
Within	Dugas et al., 2010	Within	0.18	-0.30	0.66	0.46
Within	Ost et al., 2001	Within	0.34	-0.15	0.82	0.17
Within	Flückiger et al., 2016	Within	0.04	-0.44	0.52	0.87
Within	Langkaas et al., 2017	Within	0.30	-0.18	0.79	0.22
Within	Whittal et al., 2005	Within	0.24	-0.26	0.75	0.35
Within	Kunze et al., 2017	Within	0.11	-0.40	0.63	0.66
Within	Markowitz et al., 2015	Within	0.28	-0.27	0.82	0.32
Within	Vaccaro et al., 2014	Within	0.02	-0.53	0.57	0.95
Within	Twohig et al., 2018	Within	0.25	-0.30	0.80	0.38
Within	Foa et al., 1991	Within	0.36	-0.20	0.92	0.20
Within	Budney et al., 2000	Within	0.76	0.17	1.35	0.01
Within	Borkovec et al., 1993	Within	0.33	-0.27	0.93	0.28
Within	Wells et al., 2015	Within	0.74	0.13	1.36	0.02
Within	Emmelkamp et al., 1988	Within	0.17	-0.46	0.79	0.60
Within	Goldman et al., 2007	Within	0.65	0.01	1.28	0.05
Within	Belloch et al., 2008	Within	0.18	-0.48	0.85	0.59
Within	Greenberg et al., 1998	Within	0.58	-0.09	1.25	0.09
Within	Simpson et al., 2010	Within	0.48	-0.23	1.19	0.18
Within	Taylor et al., 2003	Within	0.63	-0.08	1.35	0.08
Within	Hemanny et al., 2020	Within	0.31	-0.43	1.05	0.41
Within	Koszycki et al., 2010	Within	0.02	-0.79	0.82	0.97
Within	Emmelkamp et al., 1991	Within	0.52	-0.31	1.36	0.22
Within	Hellstrom et al., 1996	Within	0.10	-0.74	0.94	0.82
Within	Bryant et al., 2013	Within	0.11	-0.81	1.04	0.81
Within	Bryant et al., 2008	Within	0.53	-0.44	1.51	0.28
Within	Winston et al., 1994	Within	0.46	-0.54	1.46	0.37
Within	Clark et al., 1998	Within	0.15	-0.93	1.23	0.79
Within	Ost, 1988	Within	1.64	0.55	2.73	0.00
Within	Clark et al., 2006	Within	0.64	-0.45	1.74	0.25
Within	Clark et al., 1994	Within	0.73	-0.39	1.85	0.20
Within	Buckner et al., 2019	Within	0.23	-0.90	1.35	0.69
Within	Wetherell et al., 2011	Within	1.80	0.62	2.99	0.00
Within	Vincelli et al., 2003	Within	0.00	-1.21	1.21	1.00
Within	Burke et al., 1997	Within	0.42	-0.84	1.69	0.51
Within	Constantino et al., 2008	Within	0.46	-0.87	1.80	0.50
Within	Borkovec et al., 1988	Within	0.41	-0.96	1.79	0.56
Within	Pooled		0.26	0.19	0.33	0.00
Within	Prediction Interval		0.26	-0.09	0.61	
Overall	Pooled		0.30	0.24	0.36	0.00
Overall	Prediction Interval		0.30	-0.07	0.66	

Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are grouped by comparison (within the same therapy family vs. between different therapy families) and sorted in descending order of precision (i.e., ascending order of variance). The red lozenges represents the summary effect sizes for within-family comparisons, between-family comparisons, and the overall summary effect; their widths represents their 95% confidence ranges. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. All effect sizes are absolute values (i.e., given positive signs); the forest plot therefore begins with zero as the minimum value and does not depict negative values for the confidence or prediction intervals. Outlying effects were not included.

Figure J4

Forest Plot of Absolute Valued Effect Sizes of Therapy A vs. Therapy B, Global Measures at Termination

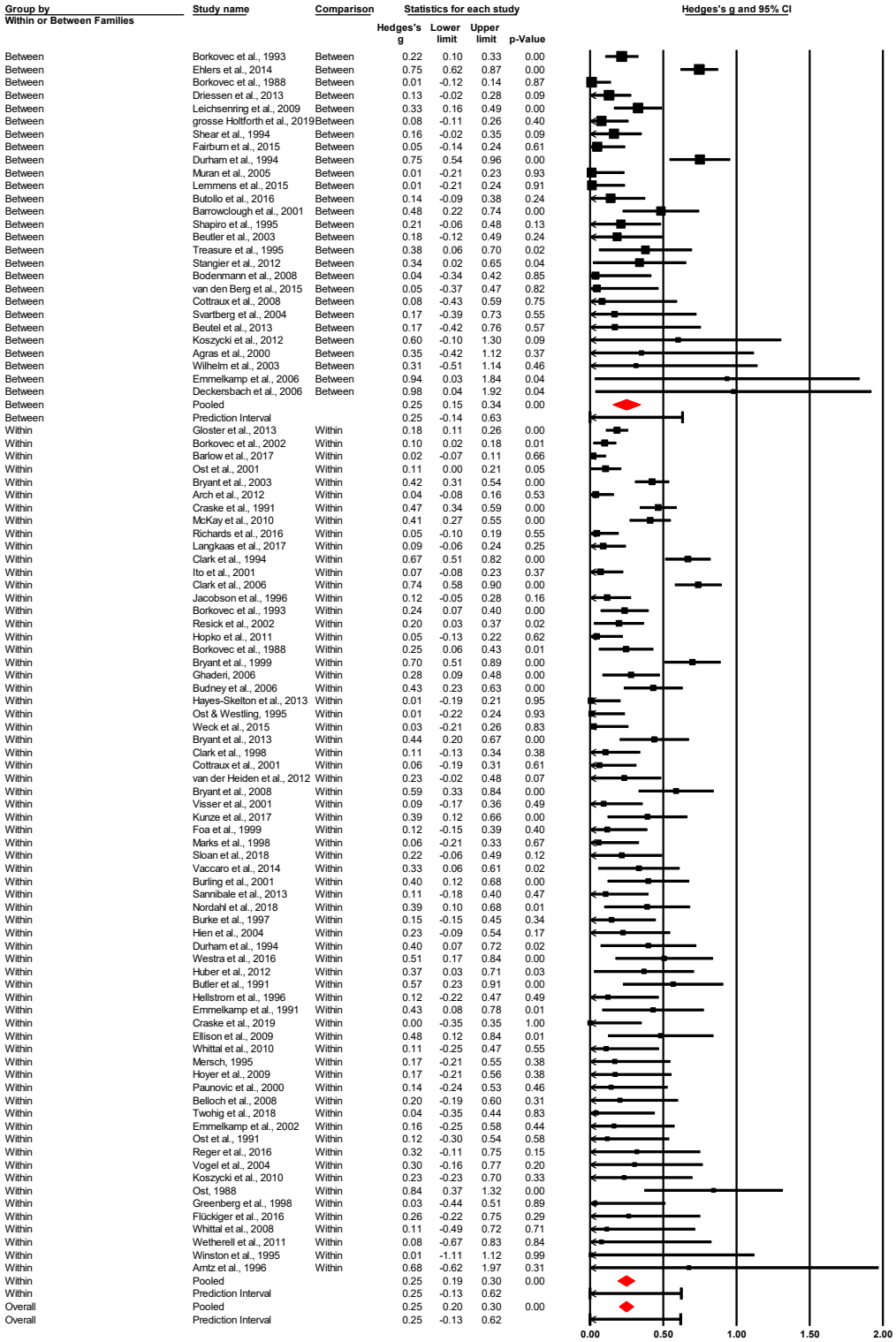
Group by Within or Between Families	Study name	Comparison	Hedges's g	Lower limit	Upper limit	p-Value
Between	Bruijniks et al., 2020	Between	0.07	-0.05	0.20	0.25
Between	Leichsenring et al., 2013	Between	0.31	0.12	0.51	0.00
Between	Kocsis et al., 2009	Between	0.15	-0.07	0.36	0.18
Between	Schramm et al., 2017	Between	0.28	0.02	0.54	0.03
Between	Lemmens et al., 2015	Between	0.02	-0.30	0.33	0.92
Between	grosse Holtforth et al., 2019	Between	0.24	-0.11	0.59	0.18
Between	Giesen-Bloo et al., 2006	Between	0.76	0.33	1.20	0.00
Between	Watson et al., 2003	Between	0.17	-0.26	0.61	0.44
Between	Vos et al., 2012	Between	1.32	0.87	1.77	0.00
Between	Horst et al., 2017	Between	0.09	-0.40	0.59	0.72
Between	Ehlers et al., 2014	Between	0.57	0.06	1.08	0.03
Between	Leichsenring et al., 2009	Between	0.02	-0.49	0.54	0.93
Between	Fonagy et al., 2020	Between	1.04	0.52	1.55	0.00
Between	Lipsitz et al., 2008	Between	0.00	-0.55	0.55	1.00
Between	Svartberg et al., 2004	Between	0.14	-0.41	0.69	0.61
Between	Muran et al., 2005	Between	0.30	-0.25	0.85	0.29
Between	Power et al., 2002	Between	0.60	0.02	1.17	0.04
Between	Shear et al., 1994	Between	0.17	-0.43	0.77	0.58
Between	Connolly Gibbons et al., 2006	Between	0.20	-0.40	0.80	0.51
Between	Markowitz et al., 2015	Between	0.20	-0.45	0.84	0.55
Between	Hersen et al., 1984	Between	0.19	-0.46	0.84	0.57
Between	Deckersbach et al., 2006	Between	0.30	-0.40	1.00	0.40
Between	Butollo et al., 2016	Between	0.20	-0.51	0.91	0.59
Between	Durham et al., 1994	Between	0.18	-0.53	0.89	0.62
Between	Koszycki et al., 2012	Between	0.75	0.04	1.46	0.04
Between	Kiosses et al., 2010	Between	0.91	0.18	1.64	0.02
Between	Hardy et al., 1995	Between	0.30	-0.43	1.04	0.42
Between	Wilhelm et al., 2003	Between	1.14	0.37	1.91	0.00
Between	Blanco et al., 2019	Between	0.12	-0.87	1.10	0.81
Between	Budney et al., 2000	Between	0.00	-1.04	1.04	1.00
Between	Beutel et al., 2013	Between	0.07	-0.97	1.12	0.89
Between	Cottraux et al., 2008	Between	0.01	-1.05	1.06	0.99
Between	Bellino et al., 2007	Between	0.57	-0.65	1.78	0.36
Between	Pooled		0.32	0.21	0.42	0.00
Between	Prediction Interval		0.32	-0.02	0.65	
Within	Richards et al., 2016	Within	0.02	-0.20	0.24	0.87
Within	Gloster et al., 2013	Within	0.59	0.36	0.82	0.00
Within	Foa et al., 2005	Within	0.07	-0.27	0.41	0.68
Within	Simpson et al., 2008	Within	0.40	-0.01	0.80	0.05
Within	Hopko et al., 2011	Within	0.12	-0.32	0.55	0.60
Within	Kramer et al., 2014	Within	0.40	-0.06	0.85	0.09
Within	Whittal et al., 2010	Within	0.16	-0.32	0.64	0.51
Within	Huber et al., 2012	Within	0.12	-0.36	0.59	0.63
Within	Langkaas et al., 2017	Within	0.23	-0.25	0.71	0.35
Within	Dugas et al., 2010	Within	0.42	-0.06	0.91	0.09
Within	Hayes-Skelton et al., 2013	Within	0.18	-0.31	0.67	0.47
Within	Jackson et al., 2018	Within	0.06	-0.44	0.56	0.80
Within	Hoyer et al., 2009	Within	0.31	-0.21	0.84	0.24
Within	Nordahl et al., 2018	Within	1.01	0.48	1.54	0.00
Within	Ghaderi, 2006	Within	0.13	-0.41	0.68	0.63
Within	Vaccaro et al., 2014	Within	0.10	-0.45	0.65	0.71
Within	Foa et al., 1999	Within	0.28	-0.33	0.88	0.37
Within	Ito et al., 2001	Within	0.09	-0.53	0.70	0.78
Within	Goldman et al., 2008	Within	0.14	-0.48	0.77	0.66
Within	Marks et al., 1998	Within	0.22	-0.41	0.84	0.50
Within	McDonagh et al., 2005	Within	0.27	-0.37	0.90	0.41
Within	Durham et al., 1994	Within	0.27	-0.38	0.91	0.42
Within	Barlow et al., 2017	Within	0.03	-0.62	0.68	0.93
Within	Mersch, 1995	Within	0.10	-0.55	0.76	0.76
Within	Greenberg et al., 1998	Within	0.59	-0.08	1.26	0.08
Within	Flückiger et al., 2016	Within	0.13	-0.55	0.82	0.70
Within	Simpson et al., 2010	Within	0.45	-0.26	1.16	0.21
Within	Arch et al., 2012	Within	0.22	-0.51	0.95	0.55
Within	Hemanny et al., 2020	Within	0.14	-0.59	0.87	0.71
Within	Markowitz et al., 2015	Within	1.06	0.30	1.82	0.01
Within	Koszycki et al., 2010	Within	0.06	-0.74	0.87	0.88
Within	Emmeikamp et al., 1988	Within	0.03	-0.85	0.91	0.95
Within	Zargar et al., 2013	Within	0.46	-0.43	1.35	0.31
Within	Cottraux et al., 2001	Within	0.12	-0.79	1.02	0.80
Within	Paunovic et al., 2000	Within	0.54	-0.41	1.49	0.27
Within	Winston et al., 1994	Within	0.55	-0.41	1.50	0.26
Within	Craske et al., 2019	Within	0.10	-0.89	1.10	0.84
Within	Clark et al., 1998	Within	0.05	-0.98	1.08	0.92
Within	Bramoweth et al., 2020	Within	0.16	-0.89	1.22	0.76
Within	Wetherell et al., 2011	Within	1.39	0.28	2.50	0.01
Within	Budney et al., 2000	Within	1.70	0.54	2.85	0.00
Within	Barlow et al., 1989	Within	0.37	-0.87	1.61	0.56
Within	Barlow et al., 1992	Within	0.63	-0.65	1.90	0.34
Within	Pooled		0.29	0.18	0.39	0.00
Within	Prediction Interval		0.29	-0.05	0.62	
Overall	Pooled		0.30	0.23	0.37	0.00
Overall	Prediction Interval		0.30	-0.01	0.62	



Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are grouped by comparison (within the same therapy family vs. between different therapy families) and sorted in descending order of precision (i.e., ascending order of variance). The red lozenges represents the summary effect sizes for within-family comparisons, between-family comparisons, and the overall summary effect; their widths represents their 95% confidence ranges. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. All effect sizes are absolute values (i.e., given positive signs); the forest plot therefore begins with zero as the minimum value and does not depict negative values for the confidence or prediction intervals. Outlying effects were not included.

Figure J5

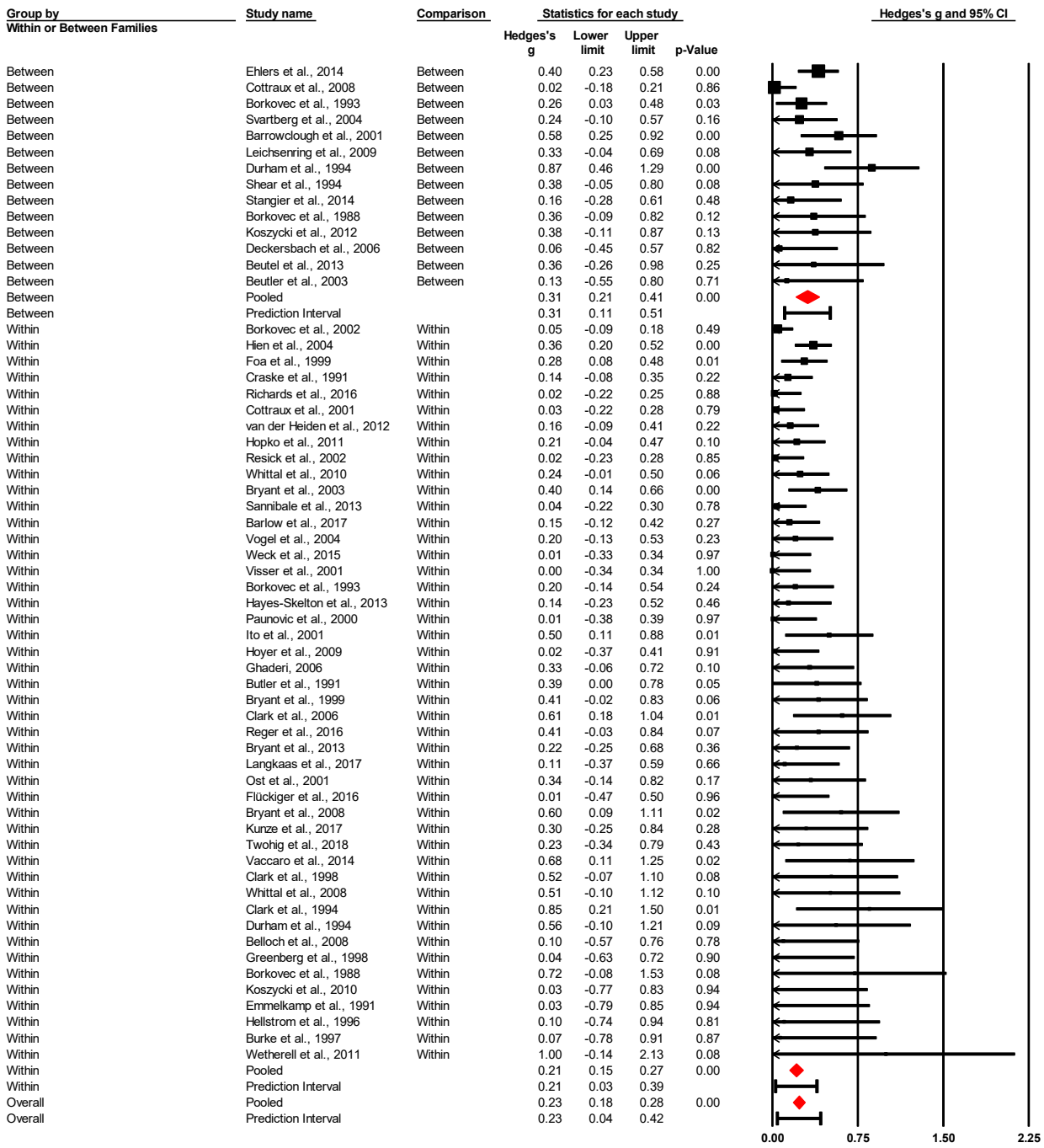
Forest Plot of Absolute Valued Effect Sizes of Therapy A vs. Therapy B, Primary Measures at Follow-Up



Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are grouped by comparison (within the same therapy family vs. between different therapy families) and sorted in descending order of precision (i.e., ascending order of variance). The red lozenges represents the summary effect sizes for within-family comparisons, between-family comparisons, and the overall summary effect; their widths represents their 95% confidence ranges. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. All effect sizes are absolute values (i.e., given positive signs); the forest plot therefore begins with zero as the minimum value and does not depict negative values for the confidence or prediction intervals. Outlying effects were not included.

Figure J6

Forest Plot of Absolute Valued Effect Sizes of Therapy A vs. Therapy B, Secondary Measures at Follow-Up



Group by Within or Between Families	Study name	Comparison	Hedges's g	Lower limit	Upper limit	p-Value
Between	Ehlers et al., 2014	Between	0.40	0.23	0.58	0.00
Between	Cottraux et al., 2008	Between	0.02	-0.18	0.21	0.86
Between	Borkovec et al., 1993	Between	0.26	0.03	0.48	0.03
Between	Svartberg et al., 2004	Between	0.24	-0.10	0.57	0.16
Between	Barrowclough et al., 2001	Between	0.58	0.25	0.92	0.00
Between	Leichsenring et al., 2009	Between	0.33	-0.04	0.69	0.08
Between	Durham et al., 1994	Between	0.87	0.46	1.29	0.00
Between	Shear et al., 1994	Between	0.38	-0.05	0.80	0.08
Between	Stangier et al., 2014	Between	0.16	-0.28	0.61	0.48
Between	Borkovec et al., 1988	Between	0.36	-0.09	0.82	0.12
Between	Koszycki et al., 2012	Between	0.38	-0.11	0.87	0.13
Between	Deckersbach et al., 2006	Between	0.06	-0.45	0.57	0.82
Between	Beutel et al., 2013	Between	0.36	-0.26	0.98	0.25
Between	Beutler et al., 2003	Between	0.13	-0.55	0.80	0.71
Between	Pooled		0.31	0.21	0.41	0.00
Between	Prediction Interval		0.31	0.11	0.51	
Within	Borkovec et al., 2002	Within	0.05	-0.09	0.18	0.49
Within	Hien et al., 2004	Within	0.36	0.20	0.52	0.00
Within	Foa et al., 1999	Within	0.28	0.08	0.48	0.01
Within	Craske et al., 1991	Within	0.14	-0.08	0.35	0.22
Within	Richards et al., 2016	Within	0.02	-0.22	0.25	0.88
Within	Cottraux et al., 2001	Within	0.03	-0.22	0.28	0.79
Within	van der Heiden et al., 2012	Within	0.16	-0.09	0.41	0.22
Within	Hopko et al., 2011	Within	0.21	-0.04	0.47	0.10
Within	Resick et al., 2002	Within	0.02	-0.23	0.28	0.85
Within	Whittal et al., 2010	Within	0.24	-0.01	0.50	0.06
Within	Bryant et al., 2003	Within	0.40	0.14	0.66	0.00
Within	Sannibale et al., 2013	Within	0.04	-0.22	0.30	0.78
Within	Barlow et al., 2017	Within	0.15	-0.12	0.42	0.27
Within	Vogel et al., 2004	Within	0.20	-0.13	0.53	0.23
Within	Weck et al., 2015	Within	0.01	-0.33	0.34	0.97
Within	Visser et al., 2001	Within	0.00	-0.34	0.34	1.00
Within	Borkovec et al., 1993	Within	0.20	-0.14	0.54	0.24
Within	Hayes-Skelton et al., 2013	Within	0.14	-0.23	0.52	0.46
Within	Paunovic et al., 2000	Within	0.01	-0.38	0.39	0.97
Within	Ito et al., 2001	Within	0.50	0.11	0.88	0.01
Within	Hoyer et al., 2009	Within	0.02	-0.37	0.41	0.91
Within	Ghaderi, 2006	Within	0.33	-0.06	0.72	0.10
Within	Butler et al., 1991	Within	0.39	0.00	0.78	0.05
Within	Bryant et al., 1999	Within	0.41	-0.02	0.83	0.06
Within	Clark et al., 2006	Within	0.61	0.18	1.04	0.01
Within	Reger et al., 2016	Within	0.41	-0.03	0.84	0.07
Within	Bryant et al., 2013	Within	0.22	-0.25	0.68	0.36
Within	Langkaas et al., 2017	Within	0.11	-0.37	0.59	0.66
Within	Ost et al., 2001	Within	0.34	-0.14	0.82	0.17
Within	Flückiger et al., 2016	Within	0.01	-0.47	0.50	0.96
Within	Bryant et al., 2008	Within	0.60	0.09	1.11	0.02
Within	Kunze et al., 2017	Within	0.30	-0.25	0.84	0.28
Within	Twohig et al., 2018	Within	0.23	-0.34	0.79	0.43
Within	Vaccaro et al., 2014	Within	0.68	0.11	1.25	0.02
Within	Clark et al., 1998	Within	0.52	-0.07	1.10	0.08
Within	Whittal et al., 2008	Within	0.51	-0.10	1.12	0.10
Within	Clark et al., 1994	Within	0.85	0.21	1.50	0.01
Within	Durham et al., 1994	Within	0.56	-0.10	1.21	0.09
Within	Belloch et al., 2008	Within	0.10	-0.57	0.76	0.78
Within	Greenberg et al., 1998	Within	0.04	-0.63	0.72	0.90
Within	Borkovec et al., 1988	Within	0.72	-0.08	1.53	0.08
Within	Koszycki et al., 2010	Within	0.03	-0.77	0.83	0.94
Within	Emmelkamp et al., 1991	Within	0.03	-0.79	0.85	0.94
Within	Hellstrom et al., 1996	Within	0.10	-0.74	0.94	0.81
Within	Burke et al., 1997	Within	0.07	-0.78	0.91	0.87
Within	Wetherell et al., 2011	Within	1.00	-0.14	2.13	0.08
Within	Pooled		0.21	0.15	0.27	0.00
Within	Prediction Interval		0.21	0.03	0.39	
Overall	Pooled		0.23	0.18	0.28	0.00
Overall	Prediction Interval		0.23	0.04	0.42	

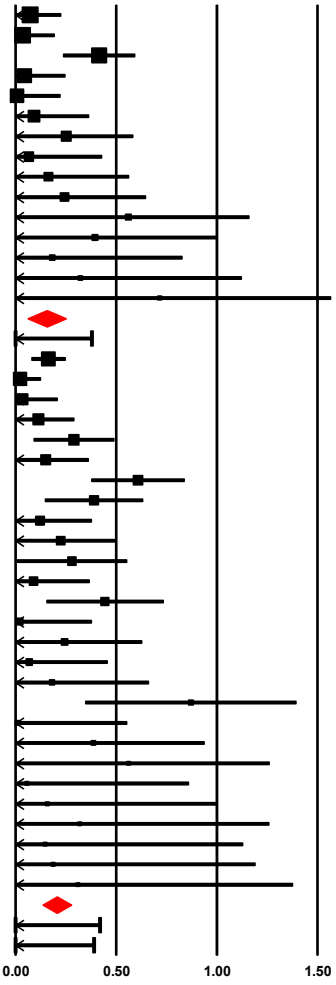
Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Horizontal bars indicate the 95%

confidence interval of each effect size. Studies are grouped by comparison (within the same therapy family vs. between different therapy families) and sorted in descending order of precision (i.e., ascending order of variance). The red lozenges represents the summary effect sizes for within-family comparisons, between-family comparisons, and the overall summary effect; their widths represents their 95% confidence ranges. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. All effect sizes are absolute values (i.e., given positive signs); the forest plot therefore begins with zero as the minimum value and does not depict negative values for the confidence or prediction intervals. Outlying effects were not included.

Figure J7

Forest Plot of Absolute Valued Effect Sizes of Therapy A vs. Therapy B, Global Measures at Follow-Up

Group by Within or Between Families	Study name	Comparison	Statistics for each study			
			Hedges's g	Lower limit	Upper limit	p-Value
Between	grosse Holtforth et al., 2019	Between	0.07	-0.08	0.23	0.36
Between	Lemmens et al., 2015	Between	0.03	-0.12	0.19	0.67
Between	Ehlers et al., 2014	Between	0.41	0.24	0.59	0.00
Between	Butollo et al., 2016	Between	0.04	-0.16	0.25	0.66
Between	Muran et al., 2005	Between	0.01	-0.21	0.22	0.95
Between	Shapiro et al., 1995	Between	0.09	-0.18	0.36	0.51
Between	Svartberg et al., 2004	Between	0.25	-0.08	0.58	0.14
Between	Leichsenring et al., 2009	Between	0.07	-0.30	0.43	0.72
Between	Koszycki et al., 2012	Between	0.16	-0.24	0.56	0.42
Between	Treasure et al., 1995	Between	0.24	-0.16	0.65	0.24
Between	Shear et al., 1994	Between	0.56	-0.04	1.16	0.07
Between	Beutel et al., 2013	Between	0.39	-0.21	1.00	0.20
Between	Cottraux et al., 2008	Between	0.18	-0.46	0.83	0.58
Between	Deckersbach et al., 2006	Between	0.32	-0.48	1.12	0.43
Between	Wilhelm et al., 2003	Between	0.72	-0.13	1.57	0.10
Between	Pooled		0.16	0.06	0.25	0.00
Between	Prediction Interval		0.16	-0.07	0.38	
Within	Hopko et al., 2011	Within	0.16	0.08	0.25	0.00
Within	Richards et al., 2016	Within	0.02	-0.08	0.13	0.67
Within	Barlow et al., 2017	Within	0.03	-0.14	0.21	0.71
Within	Arch et al., 2012	Within	0.11	-0.06	0.29	0.21
Within	Huber et al., 2012	Within	0.29	0.09	0.49	0.00
Within	Marks et al., 1998	Within	0.15	-0.06	0.36	0.17
Within	Gloster et al., 2013	Within	0.61	0.38	0.84	0.00
Within	Ellison et al., 2009	Within	0.39	0.15	0.63	0.00
Within	Hayes-Skelton et al., 2013	Within	0.12	-0.13	0.38	0.35
Within	Ito et al., 2001	Within	0.22	-0.04	0.49	0.10
Within	Foa et al., 1999	Within	0.28	0.01	0.55	0.04
Within	Langkaas et al., 2017	Within	0.09	-0.19	0.37	0.53
Within	Craske et al., 1991	Within	0.44	0.15	0.73	0.00
Within	Whittal et al., 2010	Within	0.02	-0.34	0.38	0.92
Within	Hoyer et al., 2009	Within	0.24	-0.14	0.63	0.21
Within	Vaccaro et al., 2014	Within	0.07	-0.32	0.46	0.73
Within	Greenberg et al., 1998	Within	0.18	-0.30	0.66	0.46
Within	Nordahl et al., 2018	Within	0.87	0.35	1.40	0.00
Within	Craske et al., 2019	Within	0.01	-0.53	0.55	0.97
Within	Ghaderi, 2006	Within	0.39	-0.16	0.94	0.17
Within	Flückiger et al., 2016	Within	0.56	-0.14	1.26	0.12
Within	Koszycki et al., 2010	Within	0.06	-0.75	0.86	0.89
Within	Mersch, 1995	Within	0.16	-0.68	1.00	0.71
Within	Paunovic et al., 2000	Within	0.32	-0.62	1.26	0.51
Within	Cottraux et al., 2001	Within	0.15	-0.84	1.13	0.77
Within	Clark et al., 1998	Within	0.19	-0.82	1.19	0.71
Within	Wetherell et al., 2011	Within	0.31	-0.76	1.38	0.57
Within	Pooled		0.21	0.13	0.28	0.00
Within	Prediction Interval		0.21	-0.01	0.42	
Overall	Prediction Interval		0.19	-0.01	0.39	



Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are grouped by comparison (within the same therapy family vs. between different therapy families) and sorted in descending order of precision (i.e., ascending order of variance). The red lozenges represents the summary effect sizes for within-family comparisons, between-family comparisons, and the overall summary effect; their widths represents their 95% confidence ranges. The prediction interval indicates the expected range of true effect sizes in 95% of

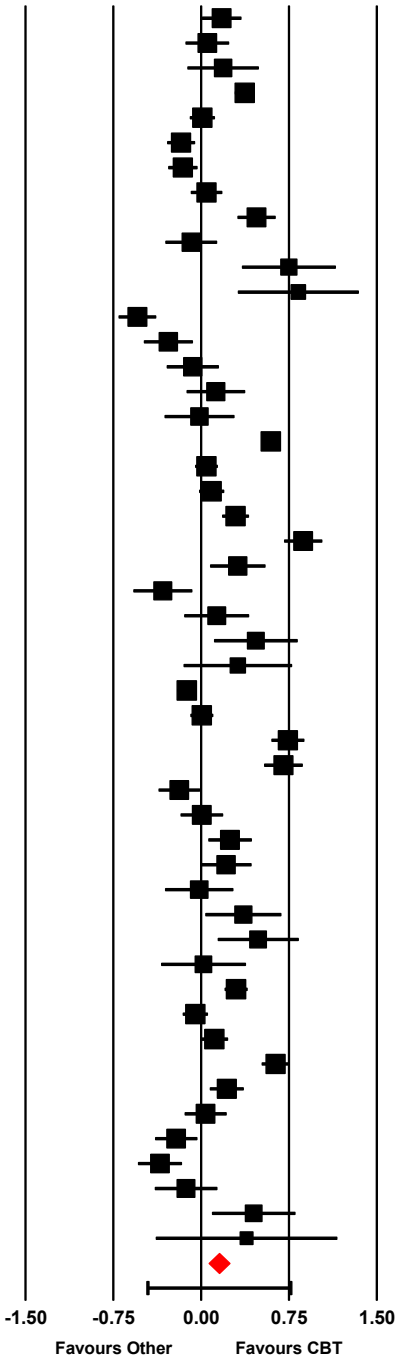
comparable populations. All effect sizes are absolute values (i.e., given positive signs); the forest plot therefore begins with zero as the minimum value and does not depict negative values for the confidence or prediction intervals. Outlying effects were not included.

Appendix K: Forest Plots of Standard Effect Sizes, CBTs vs. Other Therapies

Figure K1

Forest Plot of Effect Sizes of CBTs vs. Other Therapies, Collapsed Across Timepoints and Outcomes

Study name	Comparison	Hedges's g	Lower limit	Upper limit	p-Value
Muran et al., 2005	Combined	0.18	0.01	0.34	0.04
Blanco et al., 2019	Combined	0.05	-0.13	0.24	0.58
Beutler et al., 2003	Combined	0.19	-0.12	0.49	0.23
Borkovec et al., 1993	EHT	0.37	0.29	0.45	0.00
Borkovec et al., 1988	EHT	0.01	-0.09	0.11	0.85
Shear et al., 1994	EHT	-0.17	-0.29	-0.05	0.00
Klein et al., 2011	EHT	-0.16	-0.28	-0.03	0.01
Cottraux et al., 2008	EHT	0.05	-0.09	0.18	0.50
Barrowclough et al., 2001	EHT	0.47	0.31	0.63	0.00
Watson et al., 2003	EHT	-0.08	-0.30	0.13	0.45
Wilhelm et al., 2003	EHT	0.75	0.35	1.15	0.00
Kiosses et al., 2010	EHT	0.83	0.32	1.34	0.00
Power et al., 2002	EMDR	-0.54	-0.70	-0.39	0.00
Nijdam et al., 2012	EMDR	-0.28	-0.49	-0.07	0.01
Horst et al., 2017	EMDR	-0.07	-0.29	0.15	0.53
Taylor et al., 2003	EMDR	0.12	-0.12	0.37	0.32
van den Berg et al., 2015	EMDR	-0.01	-0.31	0.28	0.92
Ehlers et al., 2014	Integrative	0.60	0.52	0.67	0.00
grosse Holtforth et al., 2019	Integrative	0.05	-0.05	0.14	0.34
Butollo et al., 2016	Integrative	0.09	-0.02	0.20	0.09
Leichsenring et al., 2009	Integrative	0.29	0.18	0.41	0.00
Budney et al., 2000	Integrative	0.87	0.71	1.03	0.00
Deckersbach et al., 2006	Integrative	0.31	0.08	0.55	0.01
Treasure et al., 1995	Integrative	-0.33	-0.58	-0.08	0.01
Carter et al., 2013	Integrative	0.13	-0.14	0.41	0.34
Emmelkamp et al., 2006	Integrative	0.47	0.11	0.82	0.01
Nixon, 2012	Integrative	0.31	-0.15	0.77	0.18
Bruijniks et al., 2020	IPT	-0.12	-0.20	-0.04	0.00
Lemmens et al., 2015	IPT	0.01	-0.09	0.10	0.90
Fairburn et al., 2015	IPT	0.74	0.60	0.88	0.00
Vos et al., 2012	IPT	0.70	0.54	0.87	0.00
Markowitz et al., 2015	IPT	-0.19	-0.36	-0.01	0.04
Bellino et al., 2007	IPT	0.01	-0.17	0.19	0.95
Stangier et al., 2011	IPT	0.25	0.06	0.43	0.01
Carter et al., 2012	IPT	0.21	-0.00	0.43	0.05
Bodenmann et al., 2008	IPT	-0.02	-0.31	0.27	0.91
Agras et al., 2000	IPT	0.36	0.04	0.68	0.03
Zuroff et al., 2007	IPT	0.49	0.14	0.83	0.01
Bernecker et al., 2016	IPT	0.02	-0.34	0.38	0.91
Leichsenring et al., 2013	PDT	0.30	0.20	0.39	0.00
Driessen et al., 2013	PDT	-0.05	-0.16	0.06	0.36
Connolly Gibbons et al., 2006	PDT	0.11	-0.00	0.23	0.05
Durham et al., 1994	PDT	0.64	0.52	0.75	0.00
Hardy et al., 1995	PDT	0.22	0.08	0.36	0.00
Crits-Christoph et al., 1999	PDT	0.04	-0.14	0.22	0.68
Svartberg et al., 2004	PDT	-0.21	-0.39	-0.03	0.02
Beutel et al., 2013	PDT	-0.35	-0.54	-0.17	0.00
Daniel et al., 2016	PDT	-0.13	-0.39	0.14	0.34
Solomonov et al., 2019	PDT	0.45	0.09	0.80	0.01
Barber et al., 2020	PDT	0.39	-0.39	1.16	0.33
Pooled		0.15	0.07	0.24	0.00
Prediction Interval		0.15	-0.45	0.76	

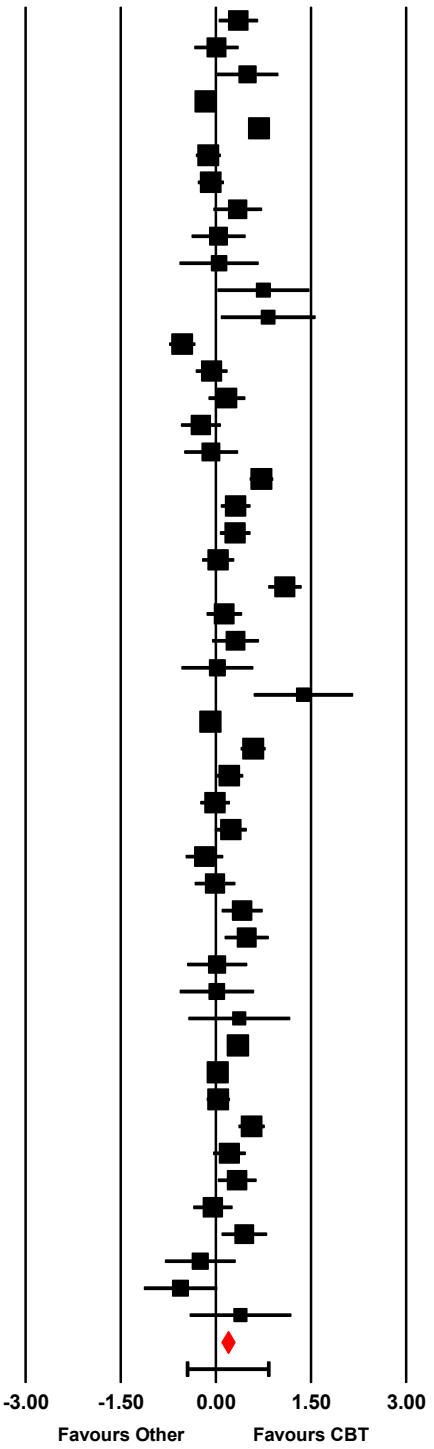


Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Positive effect sizes favour CBT. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are sorted in descending order of precision (i.e., ascending order of variance) within each comparator therapy family. The red lozenge represents the overall summary effect size across all comparator therapy families; its width represents its 95% confidence range. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. "Combined" comparisons indicate an average synthetic value, calculated to prevent dependency in the data when one study included comparisons of CBT against multiple different therapy families. Outlying effects were not included.

Figure K2

Forest Plot of Effect Sizes of CBTs vs. Other Therapies, Primary Measures at Termination

Study name	Comparison	Hedges's g	Lower limit	Upper limit	p-Value
Muran et al., 2005	Combined	0.35	0.05	0.66	0.02
Blanco et al., 2019	Combined	0.01	-0.34	0.36	0.96
Beutler et al., 2003	Combined	0.50	0.02	0.98	0.04
Klein et al., 2011	EHT	-0.16	-0.31	-0.01	0.04
Borkovec et al., 1993	EHT	0.68	0.52	0.84	0.00
Shear et al., 1994	EHT	-0.12	-0.31	0.07	0.22
Borkovec et al., 1988	EHT	-0.08	-0.28	0.12	0.43
Barrowclough et al., 2001	EHT	0.34	-0.03	0.72	0.07
Watson et al., 2003	EHT	0.04	-0.38	0.47	0.84
Cottraux et al., 2008	EHT	0.05	-0.57	0.67	0.87
Kiosses et al., 2010	EHT	0.75	0.03	1.47	0.04
Wilhelm et al., 2003	EHT	0.82	0.08	1.57	0.03
Power et al., 2002	EMDR	-0.53	-0.73	-0.33	0.00
Horst et al., 2017	EMDR	-0.07	-0.31	0.18	0.60
Taylor et al., 2003	EMDR	0.17	-0.11	0.46	0.23
Nijdam et al., 2012	EMDR	-0.24	-0.55	0.08	0.14
van den Berg et al., 2015	EMDR	-0.08	-0.50	0.34	0.72
Ehlers et al., 2014	Integrative	0.72	0.54	0.90	0.00
Leichsenring et al., 2009	Integrative	0.31	0.08	0.54	0.01
Butollo et al., 2016	Integrative	0.30	0.07	0.54	0.01
grosse Holtforth et al., 2019	Integrative	0.03	-0.21	0.28	0.78
Budney et al., 2000	Integrative	1.09	0.83	1.35	0.00
Carter et al., 2013	Integrative	0.13	-0.14	0.41	0.34
Emmelkamp et al., 2006	Integrative	0.31	-0.05	0.68	0.10
Nixon, 2012	Integrative	0.02	-0.54	0.59	0.93
Deckersbach et al., 2006	Integrative	1.38	0.60	2.16	0.00
Bruijniks et al., 2020	IPT	-0.09	-0.25	0.07	0.29
Vos et al., 2012	IPT	0.59	0.40	0.78	0.00
Carter et al., 2012	IPT	0.21	-0.00	0.43	0.05
Bellino et al., 2007	IPT	-0.01	-0.24	0.22	0.91
Fairburn et al., 2015	IPT	0.24	-0.01	0.48	0.06
Markowitz et al., 2015	IPT	-0.18	-0.47	0.11	0.23
Lemmens et al., 2015	IPT	-0.01	-0.33	0.30	0.94
Stangier et al., 2011	IPT	0.41	0.10	0.73	0.01
Zuroff et al., 2007	IPT	0.49	0.14	0.83	0.01
Bernecker et al., 2016	IPT	0.02	-0.45	0.49	0.93
Bodenmann et al., 2008	IPT	0.02	-0.57	0.60	0.96
Agras et al., 2000	IPT	0.37	-0.43	1.17	0.37
Leichsenring et al., 2013	PDT	0.35	0.21	0.48	0.00
Driessen et al., 2013	PDT	0.03	-0.12	0.18	0.70
Crits-Christoph et al., 1999	PDT	0.04	-0.14	0.22	0.68
Durham et al., 1994	PDT	0.56	0.36	0.77	0.00
Connolly Gibbons et al., 2006	PDT	0.21	-0.04	0.47	0.10
Hardy et al., 1995	PDT	0.34	0.03	0.64	0.03
Daniel et al., 2016	PDT	-0.05	-0.35	0.26	0.76
Solomonov et al., 2019	PDT	0.45	0.09	0.80	0.01
Svartberg et al., 2004	PDT	-0.25	-0.80	0.31	0.38
Beutel et al., 2013	PDT	-0.56	-1.13	0.02	0.06
Barber et al., 2020	PDT	0.39	-0.41	1.18	0.34
Pooled		0.19	0.09	0.29	0.00
Prediction Interval		0.19	-0.45	0.83	

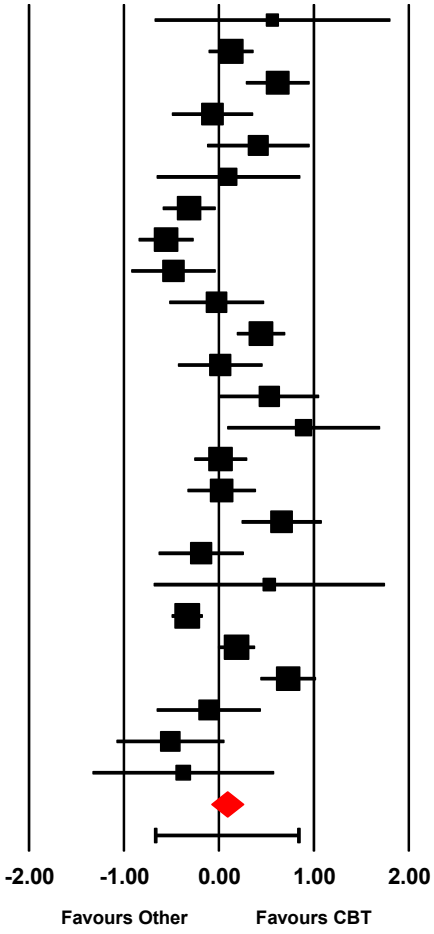


Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Positive effect sizes favour CBT. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are sorted in descending order of precision (i.e., ascending order of variance) within each comparator therapy family. The red lozenge represents the overall summary effect size across all comparator therapy families; its width represents its 95% confidence range. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. "Combined" comparisons indicate an average synthetic value, calculated to prevent dependency in the data when one study included comparisons of CBT against multiple different therapy families. Outlying effects were not included.

Figure K3

Forest Plot of Effect Sizes of CBTs vs. Other Therapies, Secondary Measures at Termination

Study name	Comparison	Hedges's g	Lower limit	Upper limit	p-Value
Beutler et al., 2003	Combined	0.56	-0.67	1.80	0.37
Cottraux et al., 2008	EHT	0.13	-0.11	0.36	0.28
Borkovec et al., 1993	EHT	0.62	0.29	0.95	0.00
Shear et al., 1994	EHT	-0.07	-0.49	0.36	0.76
Barrowclough et al., 2001	EHT	0.42	-0.12	0.95	0.13
Borkovec et al., 1988	EHT	0.10	-0.65	0.85	0.79
Nijdam et al., 2012	EMDR	-0.31	-0.59	-0.04	0.03
Power et al., 2002	EMDR	-0.56	-0.84	-0.27	0.00
Capezzani et al., 2013	EMDR	-0.48	-0.92	-0.04	0.03
Taylor et al., 2003	EMDR	-0.02	-0.52	0.47	0.92
Ehlers et al., 2014	Integrative	0.44	0.19	0.70	0.00
Deckersbach et al., 2006	Integrative	0.01	-0.43	0.46	0.95
Leichsenring et al., 2009	Integrative	0.53	0.01	1.05	0.05
Nixon, 2012	Integrative	0.89	0.09	1.70	0.03
Bruijniks et al., 2020	IPT	0.02	-0.26	0.29	0.90
Markowitz et al., 2015	IPT	0.03	-0.33	0.39	0.87
Vos et al., 2012	IPT	0.66	0.24	1.08	0.00
Stangier et al., 2013	IPT	-0.19	-0.63	0.26	0.41
Bellino et al., 2007	IPT	0.53	-0.68	1.74	0.39
Fonagy et al., 2020	PDT	-0.33	-0.49	-0.17	0.00
Leichsenring et al., 2013	PDT	0.19	-0.01	0.38	0.06
Durham et al., 1994	PDT	0.73	0.44	1.02	0.00
Svartberg et al., 2004	PDT	-0.11	-0.65	0.44	0.70
Beutel et al., 2013	PDT	-0.51	-1.08	0.06	0.08
Daniel et al., 2016	PDT	-0.37	-1.33	0.58	0.44
Pooled		0.09	-0.08	0.26	0.30
Prediction Interval		0.09	-0.66	0.84	



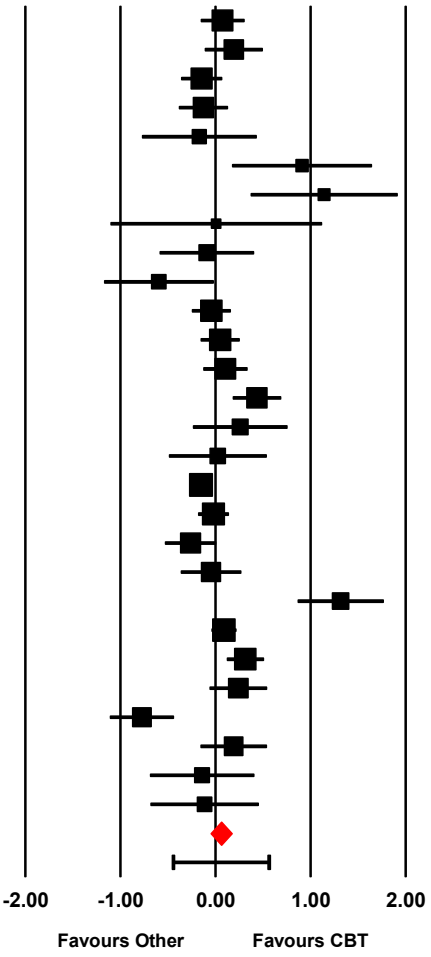
Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Positive effect sizes favour CBT. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are sorted in descending order of precision (i.e., ascending order of variance) within each comparator therapy family. The red lozenge represents the overall summary effect size across all comparator therapy families; its width represents its 95% confidence range. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. "Combined" comparisons indicate an average synthetic value,

calculated to prevent dependency in the data when one study included comparisons of CBT against multiple different therapy families. Outlying effects were not included.

Figure K4

Forest Plot of Effect Sizes of CBTs vs. Other Therapies, Global Measures at Termination

Study name	Comparison	Hedges's g	Lower limit	Upper limit	p-Value
Blanco et al., 2019	Combined	0.07	-0.15	0.30	0.52
Muran et al., 2005	Combined	0.19	-0.11	0.50	0.21
Kocsis et al., 2009	EHT	-0.15	-0.36	0.07	0.18
Watson et al., 2003	EHT	-0.13	-0.38	0.13	0.33
Shear et al., 1994	EHT	-0.17	-0.77	0.43	0.58
Kiosses et al., 2010	EHT	0.91	0.18	1.64	0.02
Wilhelm et al., 2003	EHT	1.14	0.37	1.91	0.00
Cottraux et al., 2008	EHT	0.01	-1.10	1.12	0.99
Horst et al., 2017	EMDR	-0.09	-0.59	0.40	0.72
Power et al., 2002	EMDR	-0.60	-1.17	-0.02	0.04
grosse Holtforth et al., 2019	Integrative	-0.04	-0.25	0.16	0.67
Butollo et al., 2016	Integrative	0.05	-0.15	0.25	0.63
Budney et al., 2000	Integrative	0.10	-0.13	0.33	0.38
Ehlers et al., 2014	Integrative	0.44	0.18	0.69	0.00
Deckersbach et al., 2006	Integrative	0.26	-0.24	0.75	0.30
Leichsenring et al., 2009	Integrative	0.02	-0.49	0.54	0.93
Bruijniks et al., 2020	IPT	-0.15	-0.25	-0.06	0.00
Lemmens et al., 2015	IPT	-0.02	-0.18	0.14	0.79
Markowitz et al., 2015	IPT	-0.26	-0.53	0.00	0.05
Bellino et al., 2007	IPT	-0.05	-0.36	0.27	0.77
Vos et al., 2012	IPT	1.32	0.87	1.77	0.00
Connolly Gibbons et al., 2006	PDT	0.09	-0.04	0.22	0.18
Leichsenring et al., 2013	PDT	0.31	0.12	0.51	0.00
Hardy et al., 1995	PDT	0.24	-0.06	0.54	0.12
Fonagy et al., 2020	PDT	-0.77	-1.11	-0.44	0.00
Durham et al., 1994	PDT	0.19	-0.16	0.54	0.28
Svartberg et al., 2004	PDT	-0.14	-0.69	0.41	0.61
Beutel et al., 2013	PDT	-0.11	-0.68	0.45	0.69
Pooled		0.06	-0.05	0.17	0.28
Prediction Interval		0.06	-0.44	0.56	



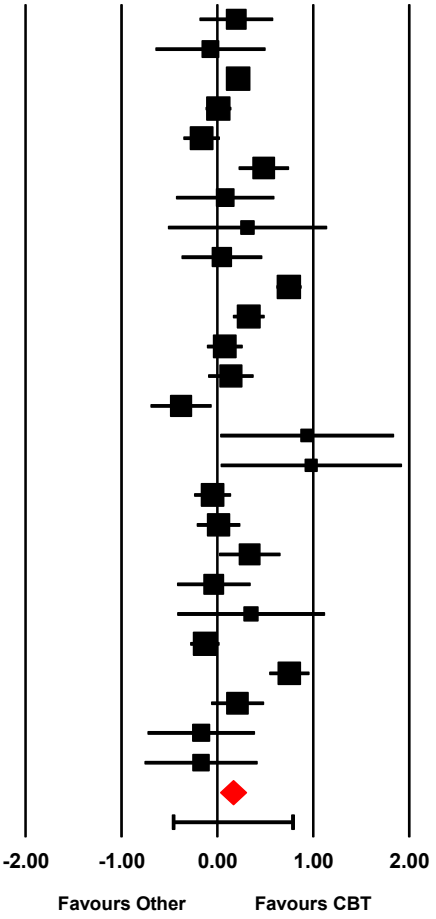
Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Positive effect sizes favour CBT. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are sorted in descending order of precision (i.e., ascending order of variance) within each comparator therapy family. The red lozenge represents the overall summary effect size across all comparator therapy families; its width represents its 95% confidence range. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. "Combined" comparisons indicate an average synthetic value,

calculated to prevent dependency in the data when one study included comparisons of CBT against multiple different therapy families. Outlying effects were not included.

Figure K5

Forest Plot of Effect Sizes of CBTs vs. Other Therapies, Primary Measures at Follow-Up

Study name	Comparison	Hedges's g	Lower limit	Upper limit	p-Value
Muran et al., 2005	Combined	0.20	-0.18	0.58	0.31
Beutler et al., 2003	Combined	-0.07	-0.64	0.50	0.81
Borkovec et al., 1993	EHT	0.22	0.10	0.33	0.00
Borkovec et al., 1988	EHT	0.01	-0.12	0.14	0.87
Shear et al., 1994	EHT	-0.16	-0.35	0.02	0.09
Barrowclough et al., 2001	EHT	0.48	0.22	0.74	0.00
Cottraux et al., 2008	EHT	0.08	-0.43	0.59	0.75
Wilhelm et al., 2003	EHT	0.31	-0.51	1.14	0.46
van den Berg et al., 2015	EMDR	0.05	-0.37	0.47	0.82
Ehlers et al., 2014	Integrative	0.75	0.62	0.87	0.00
Leichsenring et al., 2009	Integrative	0.33	0.16	0.49	0.00
grosse Holtforth et al., 2019	Integrative	0.08	-0.11	0.26	0.40
Butollo et al., 2016	Integrative	0.14	-0.09	0.38	0.24
Treasure et al., 1995	Integrative	-0.38	-0.70	-0.06	0.02
Emmelkamp et al., 2006	Integrative	0.94	0.03	1.84	0.04
Deckersbach et al., 2006	Integrative	0.98	0.04	1.92	0.04
Fairburn et al., 2015	IPT	-0.05	-0.24	0.14	0.61
Lemmens et al., 2015	IPT	0.01	-0.21	0.24	0.91
Stangier et al., 2012	IPT	0.34	0.02	0.65	0.04
Bodenmann et al., 2008	IPT	-0.04	-0.42	0.34	0.85
Agras et al., 2000	IPT	0.35	-0.42	1.12	0.37
Driessen et al., 2013	PDT	-0.13	-0.28	0.02	0.09
Durham et al., 1994	PDT	0.75	0.54	0.96	0.00
Shapiro et al., 1995	PDT	0.21	-0.06	0.48	0.13
Svartberg et al., 2004	PDT	-0.17	-0.73	0.39	0.55
Beutel et al., 2013	PDT	-0.17	-0.76	0.42	0.57
Pooled		0.17	0.03	0.30	0.02
Prediction Interval		0.17	-0.46	0.79	

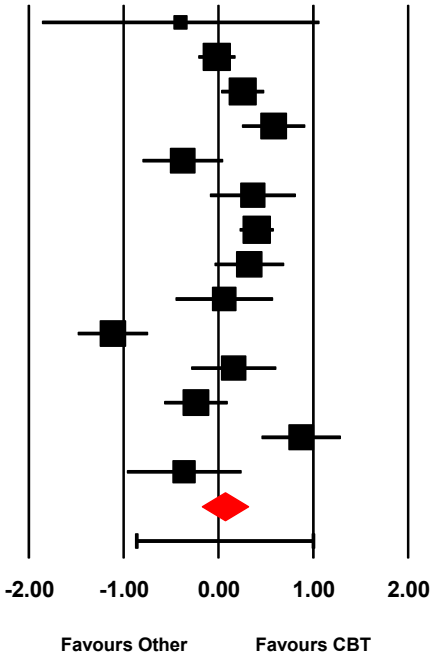


Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Positive effect sizes favour CBT. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are sorted in descending order of precision (i.e., ascending order of variance) within each comparator therapy family. The red lozenge represents the overall summary effect size across all comparator therapy families; its width represents its 95% confidence range. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. "Combined" comparisons indicate an average synthetic value, calculated to prevent dependency in the data when one study included comparisons of CBT against multiple different therapy families. Outlying effects were not included.

Figure K6

Forest Plot of Effect Sizes of CBTs vs. Other Therapies, Secondary Measures at Follow-Up

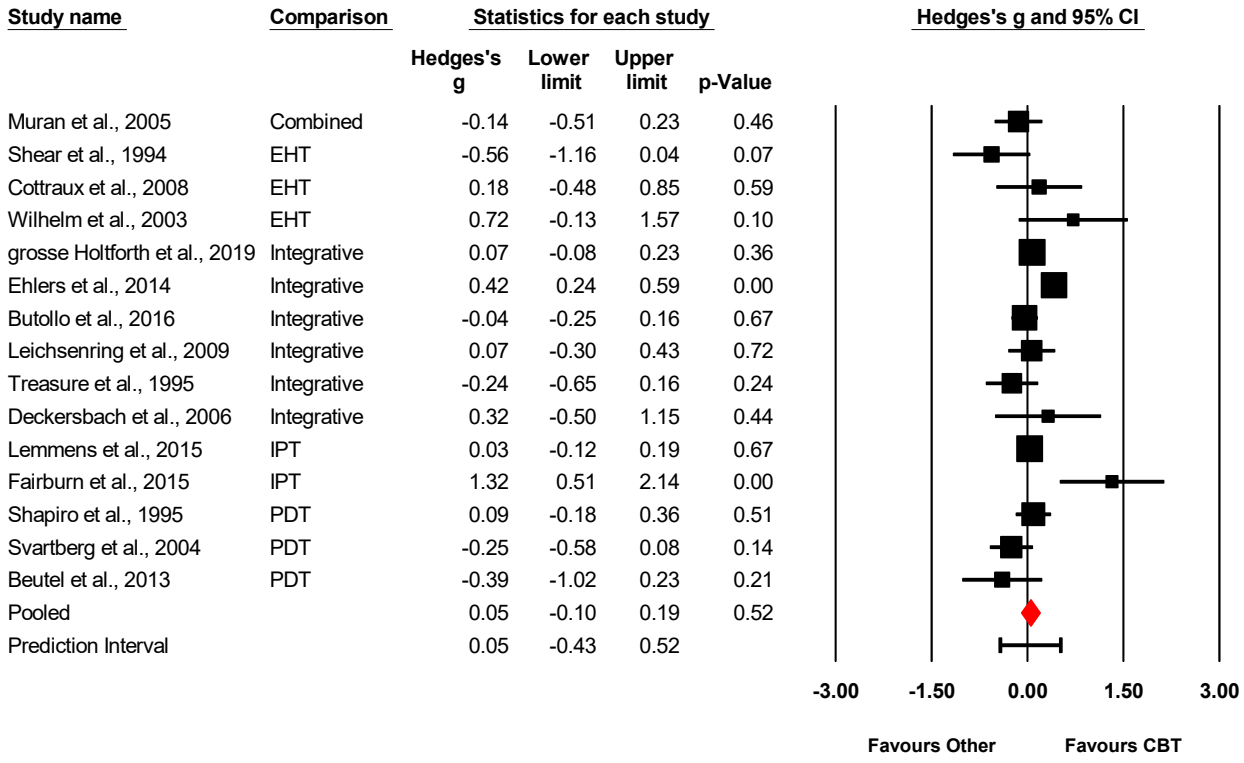
Study name	Comparison	Hedges's g	Lower limit	Upper limit	p-Value
Beutler et al., 2003	Combined	-0.40	-1.86	1.06	0.59
Cottraux et al., 2008	EHT	-0.02	-0.21	0.18	0.86
Borkovec et al., 1993	EHT	0.26	0.03	0.48	0.03
Barrowclough et al., 2001	EHT	0.58	0.25	0.91	0.00
Shear et al., 1994	EHT	-0.38	-0.80	0.05	0.08
Borkovec et al., 1988	EHT	0.36	-0.09	0.81	0.11
Ehlers et al., 2014	Integrative	0.40	0.23	0.58	0.00
Leichsenring et al., 2009	Integrative	0.33	-0.04	0.69	0.08
Deckersbach et al., 2006	Integrative	0.06	-0.45	0.57	0.82
Fairburn et al., 2015	IPT	-1.11	-1.48	-0.75	0.00
Stangier et al., 2014	IPT	0.16	-0.28	0.61	0.48
Svartberg et al., 2004	PDT	-0.24	-0.57	0.10	0.16
Durham et al., 1994	PDT	0.87	0.46	1.29	0.00
Beutel et al., 2013	PDT	-0.36	-0.96	0.24	0.24
Pooled		0.07	-0.17	0.31	0.57
Prediction Interval		0.07	-0.86	1.00	



Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Positive effect sizes favour CBT. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are sorted in descending order of precision (i.e., ascending order of variance) within each comparator therapy family. The red lozenge represents the overall summary effect size across all comparator therapy families; its width represents its 95% confidence range. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. "Combined" comparisons indicate an average synthetic value, calculated to prevent dependency in the data when one study included comparisons of CBT against multiple different therapy families. Outlying effects were not included.

Figure K7

Forest Plot of Effect Sizes of CBTs vs. Other Therapies, Global Measures at Follow-Up



Note. Each square represents the effect size of an individual comparison (Hedges's *g*), with size proportional to the weighting of the comparison in the meta-analysis. Positive effect sizes favour CBT. Horizontal bars indicate the 95% confidence interval of each effect size. Studies are sorted in descending order of precision (i.e., ascending order of variance) within each comparator therapy family. The red lozenge represents the overall summary effect size across all comparator therapy families; its width represents its 95% confidence range. The prediction interval indicates the expected range of true effect sizes in 95% of comparable populations. "Combined" comparisons indicate an average synthetic value, calculated to prevent dependency in the data when one study included comparisons of CBT against multiple different therapy families. Outlying effects were not included.

Appendix L: PRISMA Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	103
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	3
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	100
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	103
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	107, 109
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	107
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	107
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	109
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	109
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	113
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	113
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	117
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	123

Section and Topic	Item #	Checklist item	Location where item is reported
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	109
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	123, 124
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	Appendix J, Appendix K
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	123
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	116, 126
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	138, 149, 153
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	117
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	n/r
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	127
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	n/r
Study characteristics	17	Cite each included study and present its characteristics.	Appendix H
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Appendix D
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	131–164, Appendix J, Appendix K
Results of	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	127, 167

Section and Topic	Item #	Checklist item	Location where item is reported
syntheses	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	128-160
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	128-160
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	128-160
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	151
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	n/r
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	161-171
	23b	Discuss any limitations of the evidence included in the review.	161-171
	23c	Discuss any limitations of the review processes used.	161-171
	23d	Discuss implications of the results for practice, policy, and future research.	161-171
OTHER INFORMATION			
Registration and protocol	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	170
	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	171
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	n/r
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	n/r
Competing interests	26	Declare any competing interests of review authors.	n/r
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	n/r

From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71. This work is licensed under CC BY 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

Appendix M: Analyses of the Impact of Potentially Missing Recent Studies

Table M1

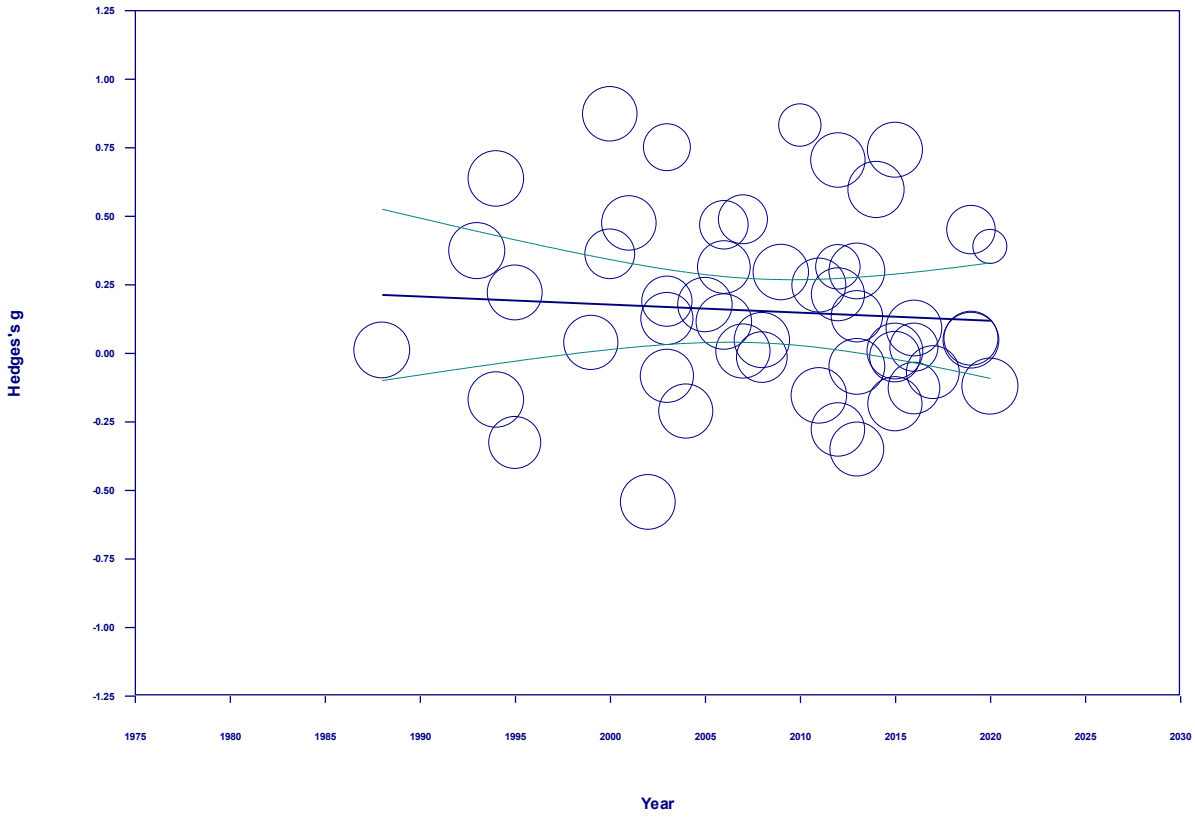
Meta-Regression of Year vs. Effect Size (Cognitive Contrast)

Analysis	All years		2015-2021	
	Coefficient	*p*-value	Coefficient	*p*-value
Timepoints and outcomes collapsed	-0.003	.61	-0.005	.89
Termination – primary	-0.008	.21	0.004	.91
Termination – secondary	-0.01	.14	-0.03*	.67*
Termination – global	-0.008	.43	-0.04*	.57*
Follow-up – primary	< -0.001	.98	0.007*	.84*
Follow-up – secondary	-0.02	.17	Insufficient data	Insufficient data
Follow-up – global	0.01	.22	0.04*	.95*

Note. Coefficient is in units of Hedges's *g*. Asterisk indicates underpowered analyses (number of studies <10).

Figure M1

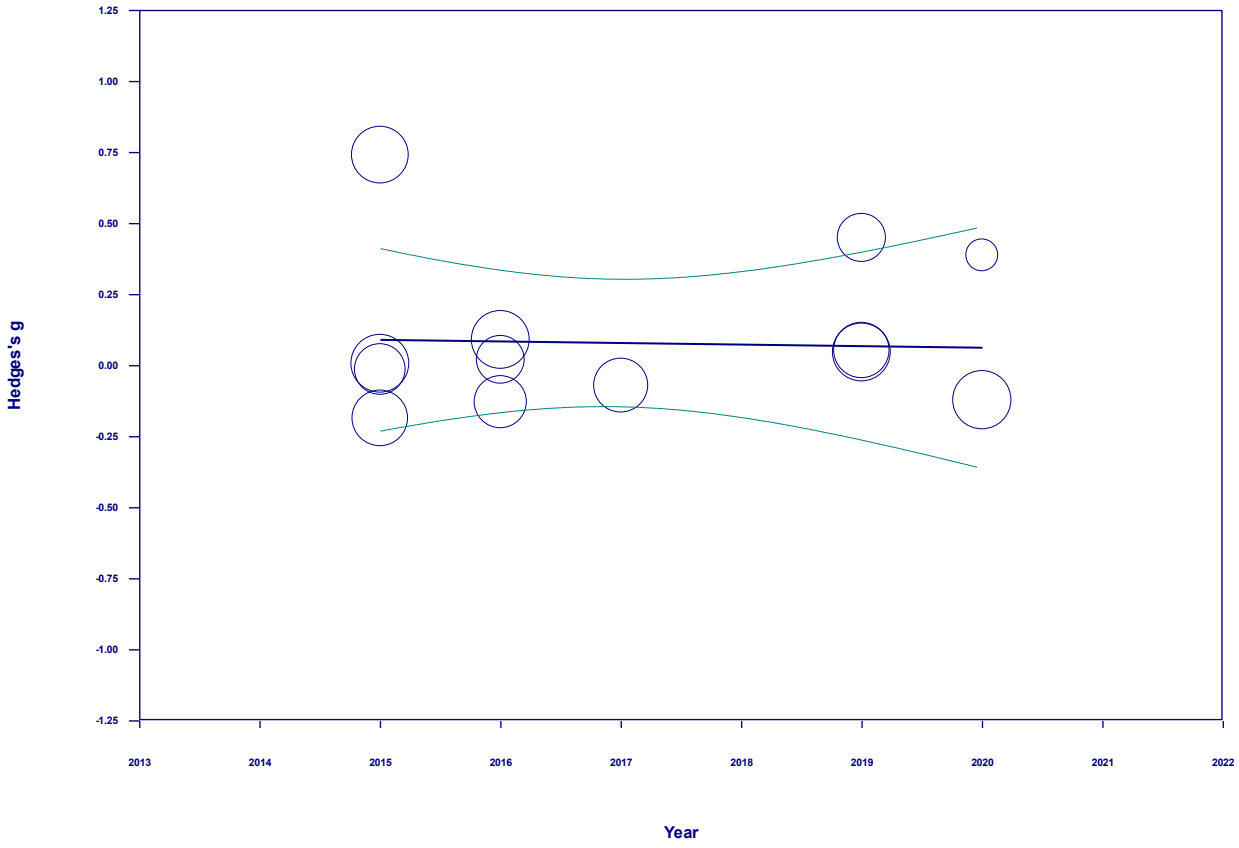
Meta-Regression of Hedge's g on Year, 1980-2021



Note. Regression of effect size against year of publication in the Cognitive Contrast analysis, collapsing across timepoints and outcomes. Each dot represents a primary RCT and is proportional to the sample size of the study. The regression line is depicted in blue. Green lines represent the bounds of the 95% confidence interval.

Figure M2

Meta-Regression of Hedge's g on Year, 2015-2021



Note. Regression of effect size against year of publication in the Cognitive Contrast analysis, collapsing across timepoints and outcomes. Each dot represents a primary RCT and is proportional to the sample size of the study. The regression line is depicted in blue. Green lines represent the bounds of the 95% confidence interval.

Table M2

Orwin's Fail-Safe N for Cognitive Contrast (k = 165).

Analysis	Observed *g* (fixed)	Studies *g* = 0 needed to dilute below threshold	
		***g* < 0.24**	***g* < 0.05, or n.s.**
Across timepoints and outcomes	0.16	n/a	113
Termination - primary	0.18	n/a	131
Termination - secondary	0.04 (n.s.)	n/a	n/a
Termination - global	0.01 (n.s.)	n/a	n/a
Follow-up - primary	0.19	n/a	76
Follow-up - secondary	0.14 (n.s.)	n/a	n/a
Follow-up - global	0.08 (n.s.)	n/a	n/a

Note. Fail-safe N analysis of studies needed to reduce observed effect size below the stated thresholds. Observed effect is the summary value of the fixed-effects analysis. All values in Hedge's *g*. "n.s." indicates "not statistically significant. "n/a" indicates the observed effect was already below the threshold value.