The Effectiveness of Psychotherapy Delivered in Routine Care Settings: A Systematic Review and Meta-Analysis

Chris Gaskell

2021-12-29

# Abstract

**Objective:** Most psychotherapy is delivered in naturalistic settings. There has been a substantial increase in recent decades in the amount of evidence generated by routine services and so this review sought to examine the effectiveness of routinely delivered psychological therapies.

**Method:** A pre-registered systematic-review and random-effects meta-analysis on studies meeting pre-specified inclusion criteria was conducted (CRD42020175235). Between-study heterogeneity was explored using methodological and clinical variables in univariate and multivariate moderator analyses.

**Results:** The systematic search identified 252 studies (*k* = 298 samples), of which 223 (*k* = 263 samples) were eligible for inclusion in the meta-analysis. Results showed large effects for depression (*d* = 0.96, [CI 0.88-1.04], *p* = < 0.001, *k* = 122), anxiety (*d* = 0.8 [CI 0.71-0.9], *p* = < 0.001, *k* = 69), and miscellaneous outcomes (*d* = 1.01 [CI 0.93-1.09], *p* = < 0.001, *k* = 158). Continent was the only significant moderator of treatment effects across all outcome domains.

**Conclusion:** This review provides support for the effectiveness of routinely delivered psychological therapy. Findings should be interpreted with caution due to the observational nature of effectiveness studies and also the marked heterogeneity shown across study designs and characteristics.

**Keywords:** ‘Psychotherapy,’ ‘Effectiveness,’ ‘Naturalistic,’ ‘Routine Outcomes,’ ‘Meta-analysis.’

**Clinical Significance of this Article:** The present study is the most comprehensive systematic review and meta-analysis of outcome data from practice-based psychotherapy studies. Pooling data from various countries, clinical contexts, therapy modalities and outcome measures, this study provides robust evidence that routinely delivered psychological interventions are effective for the treatment of depression, anxiety and global psychological distress, with large effect sizes across these domains (*d* > .80). Psychological treatments delivered in naturalistic settings are broadly effective for the alleviation of common indicators of psychological distress.

# The Effectiveness of Psychotherapy Delivered in Routine Care Settings: A Systematic Review and Meta-Analysis

# Introduction

Meta-analyses of clinical trials support the efficacy of psychotherapy for common mental health problems such as depression ([Cuijpers et al., 2008](#ref-Cuijpers2008)), anxiety disorders (e.g., [Cuijpers, Sijbrandij, et al., 2014](#ref-Cuijpers2014); [Mayo-Wilson et al., 2014](#ref-Mayo-Wilson2014); [Olatunji et al., 2014](#ref-Olatunji2014); [Sánchez-Meca et al., 2010](#ref-Sanchez-Meca2010); [Wolitzky-Taylor et al., 2008](#ref-Wolitzky-Taylor2008)), post-traumatic stress disorder ([Lewis et al., 2020](#ref-Lewis2020)), obsessive-compulsive disorder ([Rosa-Alcázar et al., 2008](#ref-Rosa-Alcazar2008)), eating disorders ([Linardon et al., 2017](#ref-Linardon2017)), psychosis ([Turner et al., 2020](#ref-Turner2020)) and other conditions. Grounded in this evidence, clinical guidelines endorse the use of psychological interventions in routine clinical care and support their inclusion in organized health care systems (e.g., [Chambless & Hollon, 1998](#ref-Chambless1998); [Chambless & Ollendick, 2001](#ref-Chambless2001); [2011](#X1365249553c325c79277ff2a4223a8026d6607c)). These guidelines typically advocate the implementation of evidence-based models of psychotherapy, closely following the procedures implemented in clinical trials and specified in treatment manuals. To this end, competency frameworks have been developed to support the dissemination of empirically supported treatments through clinical training programmes and clinical supervision (e.g., [Lemma et al., 2008](#ref-Lemma2008); [Roth et al., 2009](#ref-Roth2009); [Roth & Pilling, 2008](#ref-Roth2008)).

Despite its convincing experimental evidence base, there are reasons to expect that the effects of psychotherapy delivered in routine care settings may differ from those observed in clinical trials. Surveys of practicing clinicians reveal that many hold negative attitudes towards protocol-driven therapy and subsequently do not follow treatment manuals (e.g., [Addis & Krasnow, 2000](#ref-Addis2000)). Hence, the extent to which routinely delivered psychotherapy resembles trial-based empirically supported treatments is unclear ([Freedland et al., 2011](#ref-Freedland2011)). It has also been argued that the strict selection criteria applied in clinical trials may result in unusually homogeneous samples that do not adequately reflect clinical populations that are typical of routine care settings (e.g., [Lambert, 2013](#ref-Lambert2013a); [Zimmerman et al., 2002](#ref-Zimmerman2002)). It has also been argued that using control conditions within practice-based psychotherapy poses ethical concerns (e.g., leaving severely distressed samples untreated) and therefore various patient subgroups are likely to be under represented ([Nordmo et al., 2020](#ref-Nordmo2020); [Philips & Falkenström, 2021](#ref-Philips2021)). Differences in the clinical profiles of patients included and excluded from psychotherapy trials have been demonstrated (e.g., [van der Lem et al., 2012](#ref-vanderLem2012a)), some studies have nevertheless found similar clinical outcomes when comparing samples from efficacy trials and routine practice ([Lutz et al., 2016](#ref-Lutz2016a); [Persons et al., 1999](#ref-Persons1999)). However, this adequate transportability and effectiveness of evidence-based interventions is unlikely to be a general rule across all routine care settings. For example, significant variability in clinical outcomes has been found between clinics that implemented protocol-driven and highly standardized interventions in England ([Gyani et al., 2013](#ref-Gyani2013)). This variability may be partly explained by differences in clinical and demographic features of local populations (i.e., case-mix) and partly explained by implementation degrees of freedom: differences in how clinics manage treatment selection and treatment duration ([Clark et al., 2018](#ref-Clark2018)). For all of these reasons, it is plausible to assume that the effects of routinely delivered therapy may vary across settings and clinical populations, and may not necessarily conform to benchmarks from efficacy trials.

Given the above sources of uncertainty and variability, psychotherapy researchers have taken interest in evaluating the effectiveness of routinely delivered psychotherapy. A tradition of practice-based evidence (PBE, [Margison et al., 2000](#ref-Margison2000)) has emerged in recent decades, with numerous studies that examine the effects of routine care psychological interventions in various settings. Narrative reviews of PBE generally confirm that moderate-to-large uncontrolled (pre-to-post treatment) effect sizes are observed in routine care settings, supporting the effectiveness of psychotherapy but also demonstrating considerable variability across patient samples, therapists and clinics (e.g., see [Barkham et al., 2010](#ref-Barkham2010b); [Castonguay et al., 2013](#ref-Castonguay2013), [2021](#ref-Castonguay2021)). An inherent limitation of such narrative reviews is that they perform a selective rather than systematic and comprehensive synthesis of available data. Benchmarking studies that pool practice-based data across multiple clinics tend to report favourable pooled effects sizes, but also variability in effects across clinics (e.g., [Barkham et al., 2001](#ref-Barkham2001a); [Connell et al., 2007](#ref-Connell2007a); [Delgadillo et al., 2014](#ref-Delgadillo2014a); [Gyani et al., 2013](#ref-Gyani2013)). Although these studies help to quantify the expected magnitude of treatment effects observed in ordinary clinical settings, most are nevertheless circumscribed to small sets of clinics or geographical areas, offering limited insights into possible sources of heterogeneity in treatment outcomes. As such, systematic reviews and meta-analyses may be most illuminating.

Some meta-analytic investigations have reported that outcomes of clinic-based studies were not as favourable as those based in research settings ([Weisz et al., 1995](#ref-Weisz1995)). Other meta-analyses suggest that there are no differences in treatment effects when comparing clinic-based and efficacy studies when case-mix differences are accounted for (e.g., [Shadish et al., 1997](#ref-Shadish1997a), [2000](#ref-Shadish2000)); however, many of the included clinic-based studies applied stringent controls on the treatment procedures – making them more akin to lab-based studies. Later reviews have attempted to systematically identify available PBE studies and to derive generalizable information on the general effectiveness of psychotherapy in routine care. [Hunsley & Lee](#ref-Hunsley2007) ([2007](#ref-Hunsley2007)) reviewed 35 studies and concluded that the completion and improvement rates observed in PBE studies were comparable to those from efficacy trials. [Cahill et al.](#ref-Cahill2010) ([2010](#ref-Cahill2010)) reviewed 31 studies, concluding that psychotherapy was most effective for the treatment of common mental disorders with a pooled uncontrolled effect size of *d* = 1.29. More recently, [Wakefield et al.](#ref-Wakefield2021) ([2021](#ref-Wakefield2021)) reviewed 60 studies, of which 47 were eligible for meta-analysis. They reported large uncontrolled effect sizes for depression (*d* = 0.87) and anxiety (*d* = 0.88), and a moderate effect on functional impairment (*d* = 0.55).

Heterogeneous effects are commonly found in psychotherapy meta-analyses, and can, to some extent, be explained through potential moderator variables ([Spielmans & Flückiger, 2018](#ref-Spielmans2018)). More recent PBE meta-analyses have started to provide insights into plausible methodological (e.g., completers analyses vs. inclusion of patients lost to follow-up) and clinical sources of heterogeneity in effectiveness (e.g., larger effects for common mental disorders, lower effects for patients with comorbidities and socioeconomic disadvantages, larger effects for lengthier interventions). Nevertheless, these meta-analyses are over a decade old ([Cahill et al., 2010](#ref-Cahill2010); [Hunsley & Lee, 2007](#ref-Hunsley2007)) or limited to a specific primary care setting ([Wakefield et al., 2021](#ref-Wakefield2021)). Further research into the methodological and clinical sources of treatment heterogeneity is needed for a more complete understanding of underlying differences in how people respond to treatment ([Spielmans & Flückiger, 2018](#ref-Spielmans2018)).

Recent meta-analyses of efficacy trials have provided evidence for a range of potential moderators variables that are yet to be confirmed through meta-analysis of PBE. Methodological risk of bias criteria have often been associated with overestimates of treatment effect (for a review see [Munder & Barth, 2018](#ref-Munder2018)). [Barth et al.](#ref-Barth2013a) ([2013](#ref-Barth2013a)) for example observed inflated effect sizes for psychotherapy treatment studies of depression that: (i) lacked concealment of randomization, (ii) used non-blinded outcome assessors, and (iii) employed small samples. [Cuijpers et al.](#ref-Cuijpers2010) ([2010](#ref-Cuijpers2010)) also found that exclusion of patients lost to follow-up (i.e., completer samples) produced inflated effect-sizes over studies that included such patients (i.e., intention-to-treat); although [Barth et al.](#ref-Barth2013a) ([2013](#ref-Barth2013a)) found conflicting evidence.

Various clinical factors have been considered for their potential to explain heterogeneity of psychotherapy treatment findings (see [Spielmans & Flückiger, 2018](#ref-Spielmans2018) for a review). For treatment modality (e.g., cognitive-behavioural, psychodynamic) recent meta-analyses of RCTs directly comparing different treatments have provided limited evidence that there are significant differences between treatments for depression ([Barth et al., 2013](#ref-Barth2013a); [Cuijpers et al., 2008](#ref-Cuijpers2008)). Similar investigations in other conditions (e.g., generalised anxiety) have been restricted by the lack of evidence from direct comparison trials ([Cuijpers, Sijbrandij, et al., 2014](#ref-Cuijpers2014)). Aggregation of evidence for specific treatment modalities from non-comparative studies, although observational and highly susceptible to confounds, can allow for indirect comparisons of treatments.

Treatment dosage (or duration) is a clinically and economically relevant treatment variable however is often absent from psychotherapy meta-analytic reviews ([Spielmans & Flückiger, 2018](#ref-Spielmans2018)). While there are various dosage prototypes available (see [Flückiger et al., 2020](#ref-Fluckiger2020)) it is generally recommended that meta-analytic reviewers code it dosage as a continuous variable ([Spielmans & Flückiger, 2018](#ref-Spielmans2018)). Several recent meta-analyses have failed to demonstrate a significant differential effect for dosage in depression (e.g., [Barth et al., 2013](#ref-Barth2013a); [Cuijpers et al., 2013](#ref-Cuijpers2013)), anxiety (e.g., [Carpenter et al., 2018](#ref-Carpenter2018)) and OCD (e.g., [Olatunji et al., 2015](#ref-Olatunji2015)); although there remain exceptions to this rule (e.g., [Olatunji et al., 2014](#ref-Olatunji2014); [Turner et al., 2020](#ref-Turner2020)). Consideration of dosage within naturalistic settings is particularly relevant as the range of sessions is often treated more flexibly ([Lambert, 2013](#ref-Lambert2013a)) and varies between services and mental health systems ([Flückiger et al., 2020](#ref-Fluckiger2020)).

Sample severity (i.e., clinical complexity of the sample) has been widely considered to be negatively associated with treatment effect. This has received meta-analytic support through comparatively smaller pooled effect-sizes shown in meta-analyses of severe depression (e.g., [Cuijpers et al., 2010](#ref-Cuijpers2010)) and inpatient samples ([Cuijpers et al., 2011](#ref-Cuijpers2011)). Differences between stratifications of depression severity have also been shown, with more severe samples showing smaller effect sizes ([Whiston et al., 2019](#ref-Whiston2019)). Despite this, conflicting evidence have been reported through study level meta analyses of depression ([Cuijpers, Turner, et al., 2014](#ref-Cuijpers2014a); [Driessen et al., 2010](#ref-Driessen2010)), and health anxiety ([Olatunji et al., 2014](#ref-Olatunji2014)) and independent patient level meta-analyses of depression ([Furukawa et al., 2017](#ref-Furukawa2017)) and psychosis ([Turner et al., 2020](#ref-Turner2020)). [Furukawa et al.](#ref-Furukawa2017) ([2017](#ref-Furukawa2017)) for example found that patients receiving CBT can benefit across the spectrum of pre-treatment severity of depression. Varying levels of support have also been provided for differential rates of effectiveness based upon role of the outcome assessor (self-report vs clinician, [Cuijpers, Sijbrandij, et al., 2014](#ref-Cuijpers2014); [Cuijpers et al., 2008](#ref-Cuijpers2008)), sensitivity of measurement tool (e.g., [Cuijpers, Sijbrandij, et al., 2014](#ref-Cuijpers2014)), and country. The degree to which these moderator variables findings are consistent within naturalistic settings remains unclear.

The considerable growth of the PBE literature in the last decade and implementation of empirically supported treatments across many settings warrants a comprehensive review and quantitative synthesis of the literature. This would enable us to gain a more precise understanding of [1] the magnitude of effect sizes across multiple outcome domains and [2] sources of heterogeneity. The aim of the present study was to systematically review available PBE studies using a meta-analytic synthesis of quantitative data and pre-specified moderator analyses informed by earlier studies.

**Note:** The authors recognise that use of the term *effectiveness* may be somewhat misleading. The pre-post (uncontrolled) methodology which forms the body of evidence in this review is unable to disentangle treatments effects from other potential causes of change (e.g., regression to the mean, placebo). Observed change in symptoms may therefore not exclusively represent treatment effectiveness. We have opted to retain use of this term within the current review because it has frequently been used as such in the extant literature (e.g., [Lambert, 2013](#ref-Lambert2013a); [Nordmo et al., 2020](#ref-Nordmo2020)).

# Method

## Search Strategy and Eligibility

A systematic review and meta-analysis were conducted and reported using the Preferred Reporting Items for Systematic Review and Meta-Analysis guidelines (PRISMA, [Page et al., 2021](#ref-Page2021)) and the meta-analyses in psychotherapy (MAP-24, [Flückiger et al., 2018](#ref-Fluckiger2018)) guidelines. The review protocol was pre-registered through PROSPERO (CRD42020175235)[[1]](#footnote-1).

Literature was searched for articles reported prior to the search date (-April 2020) using the following inclusion criteria: (a) treatment delivered in a service which offers routine treatment (i.e., not primarily for research); (b) all adult sample (no patients under 16); (c) employed a *psychological treatment* (i.e., driven by psychological theory and intended to be therapeutic [[Spielmans & Flückiger, 2018](#ref-Spielmans2018)], as inferred or described by study manuscripts, family/group treatments were excluded); and (d) conducted face-to-face. Studies were then excluded if they: (e) were not available in English; (f) did not employ a self-report measure of treatment effectiveness (i.e. process, predictor, well-being or satisfaction measures not accepted); (g) did not provide an outcome for the acute treatment stage (i.e., pre-post comparison, or in the absence of post-treatment score the soonest available follow-up measurement to a maximum of 6-months); or (f) employed randomization procedures or control groups. Although randomised control trials seek to reduce internal bias, their conditions are often not typical of routine services as the need for patient consent will lead to bias in who is offered and who accepts these conditions, therefore reducing representativeness. For review PICO table see supplementary Table 1.

There were three phases to the systematic search. Phase one was a systematic search of three electronic literature databases (MEDLINE, CINAHL and PsycInfo) via EBSCO using a pre-developed list of key terms. Study titles/abstracts were required to have a methodologically *AND* psychologically relevant term. Methodological terms (indicating effectiveness) included: *practice-based evidence*, *routine practice*, *benchmarking*, *transportability*, *transferability*, *clinically representative*, *managed care setting*, *uncontrolled*, *external validity*, *applicable findings*, *empirically supported*, *dissemination*, and *clinical effectiveness evaluation*. These terms were selected based on their use in prior reviews of psychotherapy effectiveness ([Cahill et al., 2010](#ref-Cahill2010); [Stewart & Chambless, 2009](#ref-Stewart2009a)). *Effectiveness* and *evaluation* were not used as single word terms due to producing unmanageable numbers of hits. For the psychologically relevant term: *psycho*\* OR *therap*\* was used for PsycInfo while *psycho*\* alone was used for MEDLINE and CINAHL (*therap*\* was removed due to producing an unmanageable number of irrelevant hits). Limiters included *adult population* and *English language*. No exclusions were made based on the type of publication. Key term combinations and Boolean operators are reported in supplementary Table 2. Phase two included (a) a manual search of reference lists, and (b) forward citation searching (using Google Scholar) for all studies identified in phase one. Article titles relevant to the current review were identified by the first author using the search terms. Finally, phase three was a pragmatic grey literature search using the terms *psychotherapy* AND *routine-practice* AND *effectiveness* in GoogleScholar before reviewing the first 50 pages of results. When contacting study authors for additional information an accompanying invitation to recommend additional studies was made.

After removal of duplicates search results were imported to *Rayyan* ([Ouzzani et al., 2016](#ref-Ouzzani2016)). Rayyan is a web-platform that supports title/abstract screening through blinding decision results among collaborators and sorting abstracts by probability of inclusion through text mining. Studies identified from the systematic search were screened by the first author using a pre-developed and piloted screening tool. A sub-sample were screened by a second coder at each stage. Percentage agreement and inter-rater reliability statistics (Kappa [], [Cohen, 1960](#ref-Cohen1960)) were used to quantify screening precision. Descriptive classifiers available for interpreting were employed ([Landis & Koch, 1977](#ref-Landis1977)), consisting of *slight* (0-0.2), *fair* (0.2-0.4), *moderate* (0.4-0.6), *substantial* (0.6-0.8), and *almost perfect* (0.8-1.0). 20% of titles/abstracts were coded by a trainee clinical psychologist showing substantial reliability ( = 0.78, 1713/1740, 98.45%) and 10% of full texts were coded by a clinical psychologist, showing strong reliability ( = 0.65, 24/30, 80%). For many included studies (212/252, 84.13%) authors were contacted via e-mail for additional information (two-week response time). 177 requests were for missing correlations while 35 were for additional data (e.g. M, SD etc.). E-mail responses were received for 76 authors (35.85%) while data was provided for 41 samples (19.34%).

## Extraction

There was three separate outcome domains (and subsequently three meta-analyses) for *depression*, *anxiety* and *miscellaneous* outcomes. For anxiety and depression, domain allocation was informed by the measure employed (i.e., depression measures in *depression*, anxiety measures in *anxiety*). All other effectiveness measures were placed in the *miscellaneous* domain; which commonly consisted of (i) general psychological distress scales, and then to a lesser extent (ii) peripheral outcomes scales indicating symptom amelioration (e.g., functioning, quality of life), or (iii) diagnosis-specific outcome scales (e.g., OCD, PTSD). Samples were not exclusive to disorder specific populations/treatments and therefore domains should not be seen as synonymous with diagnosis/treatment.

A standardised extraction sheet was developed and pilot-tested with a sample of studies (*k* = 10). When extracting, pooled study samples were preferred. When only multiple independent samples were reported then effect-sizes were aggregated prior to meta-analysis to reduce bias of statistical dependency ([Gleser & Olkin, 2009](#ref-Gleser2009); [Hoyt & Del Re, 2018](#ref-Hoyt2018)). To avoid loss of information, study samples were disaggregated for moderator analyses ([Cooper, 1998](#ref-Cooper1998)). Studies with overlapping datasets were excluded. Samples which analysed patients lost to follow up were preferred to completer samples due to being less prone to attrition bias ([Jüni et al., 2001](#ref-Juni2001)). As extraction of multiple study effect-sizes within a single domain would lead to statistical dependency ([Borenstein et al., 2021](#ref-Borenstein2021)) we selected a single effect-size per sample, per domain ([Card, 2015](#ref-Card2015); [Cuijpers, 2016](#ref-Cuijpers2016)), using a preference system (defined a priori, supplementary material) that favored the most commonly employed measures in routine practice. As the current review only included self-report measures the need to select among multiple measures was rarely required. Reliability of coding for effect-size data was computed using a second coder for a sub-sample of manuscripts (n = 29) demonstrating almost perfect reliability across all values ( = 0.97, agreement = 97.56%) and perfect reliability for effect-size values ( = 1.00). Key categorical and numerical variables extracted from manuscripts for moderator analyses are reported in Table 1.

|  |
| --- |
| Table 1: Summary coding sheet for extracting study information. These moderators form the subgroup and continuous variables moderator variables for the current study. |
| Categorical variables |
| * **Setting:** the study was (i) *out-patient*, (ii) *inpatient* or (iii) *mixed.* |
| * **Analysis:** samples (i) *included* or (ii) *excluded* (completers) patients lost to follow up. |
| * **Severity:** was determined through a stratification of studies based on characteristics of the service (similar to the approach used by [de Jong et al., 2021](#ref-deJong2021a)). (i) *Mild services* included primary care, physical health, University counselling, voluntary, private (independent or group) and employee assistance programmes; (ii) *Moderate services* included secondary care, community mental health centers, specialist psychotherapy centers, managed care settings, or intensive outpatient programmes; (iii) *severe services* represented inpatient samples; and (iv) *university* included University out-patient and training clinics. |
| * **Treatment modality:** Treatments were coded as (i) *cognitive-behavioral* or (ii) *psychodynamic* based on manuscript self-designation (i.e., if the manuscript described treatment as CBT, then that was coded). In the absence of these terms, modality of best-fit was decided using treatment descriptions. Treatments that could not be confidently allocated to these groups were coded as (iii) *counseling* (e.g., person-centered, undefined) or (iv) *other*. Treatments that did not describe treatment modality were rated as other. |
| * **Continent:** Studies were coded as North America, United Kingdom (UK), mainland Europe, Australasia, or Asia. The UK was separated from Europe because of the high representation of outcomes research coming from the UK. |
| * **Intervention development stage:** Studies were coded as (i) *preliminary studies* (i.e., testing novel treatments or treatment iterations) or (ii) *routine evaluations*. |
| * **Experience:** Samples for which treatment delivery was exclusively by (i) *training professionals*, or (ii) *qualified professionals* |
| * **Measurement tool:** Measures that were represented at least ten times in the meta-analysis were entered as subgroups |
| * **Sample Size:** Following the approach of [Barth et al.](#ref-Barth2013a) ([2013](#ref-Barth2013a)), studies were coded as small (N=<25), medium (N=25-50), or large (N=50+). |
| Continuous variables |
| * **Age:** Sample mean average age. |
| * **Dosage:** Sample mean average treatment sessions. |
| * **Year:** of publication. |
| * **Female preponderance**: Sample rate (%). |
| * **Married**: Sample rate (%). |
| * **Employed**: Sample rate (%). |
| * **Ethnic minority**: Sample rate (%). |

### Risk of Bias and Quality Assessment

The Joanna Briggs Institute Quality Appraisal Tool for Case Series ([Munn et al., 2020](#ref-Munn2020)) was used to rate study risk of bias. Eight criteria primarily focusing upon manuscript reporting detail were used. Criteria included manuscript reporting of: (i) service inclusion criteria, (ii) service description, (iii) treatment description, (iv) sample characteristics, (v) outcome data, (vi) effect-size calculation, (vii) consecutive patient recruitment, and (viii) inclusion of patients lost to follow-up (in statistical analysis). Each item was coded as either met or not met (including not clear) by the first author for each sample. A sub-sample (23.8%) was second coded by a pair of MSc psychological research methods students (11.9% each).

The methodological quality for the body of evidence within each meta-analytic domain was assessed by three reviewers using guidelines for the Grading of Recommendations, Assessment, Development and Evaluations (GRADE, [Guyatt et al., 2008](#ref-Guyatt2008)). This framework rates evidence quality for each meta-analytic outcome based on included study designs. Individual ratings are initially provided (high, moderate, low or very low) and are then down-graded (or upgraded) through evaluation of five separate criteria; risk of bias within included studies, inconsistencies in aggregated treatment effect, indirectness of evidence, imprecision and publication bias.

## Analysis

All analyses were conducted using the R statistical analysis environment ([R Core Team, 2020](#ref-R-base), v 4.0.2). Reporting of effect-size calculation followed available guidance ([Hoyt & Del Re, 2018](#ref-Hoyt2018)). We calculated the standardised mean change (SMC: [Becker, 1988](#ref-Becker1988)) for included studies using the *metafor* package. This approach divides the pre-post mean change score by the pretreatment standard deviation. Calculation of the sampling variance required an estimate of the pre-post correlation ([Morris, 2008](#ref-Morris2008)). For manuscripts not reporting all required information an approach to obtaining and/or imputing was followed (designed a priori, supplementary Table 3). When unavailable, Pearson’s *r* was imputed using an empirically derived estimate (*r* = .60, [Balk et al., 2012](#ref-Balk2012)). If all steps of this approach were unsuccessful then the study was removed from the meta-analysis. Aggregation of study samples (or sampling errors) was conducted using the *aggregate* function of *metafor* using standard inverse-variance weighting.

Meta-analyses were performed using the *metafor* ([Viechtbauer, 2020](#ref-R-metafor)), *dmetar* ([Harrer et al., 2019a](#ref-Harrer2019a)), and *meta* ([Schwarzer, 2020](#ref-R-meta)) packages. Due to expected high heterogeneity, random-effects meta-analyses were used to estimate pooled and weighted effect-sizes ([Higgins & Green, 2008](#ref-Higgins2008)). This approach holds the assumption that included studies are randomly sampled from a population of studies ([Borenstein et al., 2021](#ref-Borenstein2021)). 95% confidence intervals were calculated for included studies. Forest plots were used to visualise the pattern of effects however due to the high number of studies they were stripped of text. As the large differences in study N risked disproportionate weighting of small samples ([Borenstein et al., 2021](#ref-Borenstein2021)) a post-hoc sensitivity analysis was performed through re-running the primary meta-analyses using fixed-effects models. Between-study heterogeneity was assessed using I2 ([Higgins & Thompson, 2002](#ref-Higgins2002)) and the Q statistic ([Cochran, 1954](#ref-Cochran1954)). I2 was interpreted as *low* (25-50%), *moderate* (50-75%) or *high* (75-100%, [Higgins et al., 2003](#ref-Higgins2003)). The impact of publication bias on treatment estimates was visualised using funnel plots and assessed statistically using rank correlation tests ([Begg & Mazumdar, 1994](#ref-Begg1994)), Egger’s regression test for funnel plot asymmetry ([Egger et al., 1997](#ref-Egger1997)), and fail-safe N (Rosenthal method, [Rosenthal, 1979](#ref-Rosenthal1979)).

Heterogeneity of SMC scores were tested using a range of pre-defined continuous (i.e., meta-regression) and categorical (i.e., subgroup analysis) moderator variables. Subgroup variables included: (i) *setting* (inpatient vs outpatient), (ii) *analysis* (i.e., non-/inclusion of patients lost to follow-up), (iv) *continent*, (iii) *severity* (mild, moderate, severe, university) (v) *treatment modality*, (vi) *experience* (unqualified vs. qualified therapists) (vii) *stage of treatment development* (preliminary study vs. routine evaluations), (viii) *measurement tool*, and (ix) *sample size* (small, medium, large). Meta-regression variables included (i) *publication year*, (ii) average *age* of sample, (iii) *dosage* (i.e. outpatient treatment sessions), and sample characteristics (% of samples: *female*, *minority ethnicity*, *married*, and in full-time *employment*. Moderator analyses followed available guidance ([Harrer et al., 2019b](#ref-Harrer2019)). Studies that were not relevant to a specific moderator analysis (or did not report required information) were temporarily omitted. All moderator analyses utilised mixed effect models ([Borenstein et al., 2021](#ref-Borenstein2021)) with weighted estimation (inverse-variance weights). This approach uses random effects models for subgroup pooled effect sizes and fixed-effect models for testing differences between subgroups. An omnibus test (QM-test) was used to assess for significant subgroup differences (*p* = < .05). Significant moderators were considered through inspection of 95% confidence intervals nonoverlap.

Multivariate meta regression was used to assess two models specified a priori. These included (i) *analysis*\**dosage*, and (ii) *continent*\**dosage*. Models were first developed using the subgroup (i.e., dummy) variable only, then with inclusion of dosage, and finally allowing for interactions. Significant models (*p* = <.05) were compared for goodness-of-fit using log-likelihood score, and Akaike’s-information criteria (AIC).

# Results

## Search Results

A PRISMA flow diagram ([Page et al., 2021](#ref-Page2021)) is shown in Figure 1. 10,503 records were identified through the systematic search. Following title/abstract screening there was 619 studies remaining. 30 manuscripts were not available through the first author’s institution. E-mail requests to corresponding authors led to the retrieval of 8 further manuscripts. Articles with unavailable manuscripts (n = 22) were excluded from the review. Following full-text screening there were 252 manuscripts remaining (*k* = 298), of which 223 (*k* = 263) had sufficient information to be included in the meta-analysis. Summary statistics are provided in Table 2.

### Study Characteristics

Study publication ranged from 1984 to 2020 (median = 2013, *k* = 294 published ≥ 2000). 169 samples included patients lost to follow-up (*k* = 118 completers). There were 34 (11.41%) preliminary study samples and 264 (88.59%) routine practice evaluation samples. The USA was the most well represented country (*k* = 113), followed by England (*k* = 78), Germany (*k* = 24), Sweden (*k* = 12), and Canada (*k* = 10). These five most well-represented countries accounted for the majority of the included samples (*k* = 237). For continent, North America (*k* = 123) was the most well represented, followed by the UK (*k* = 96), mainland Europe (*k* = 63), Australasia (*k* = 10), and Asia (*k* = 6).

**Figure 1:**

Prisma flow diagram of studies throughout the review.

**Identification of studies via databases and registers**

Records removed *before screening*:

Duplicate records removed (n = 1794)

Records identified from\*:

* Databases (n = 10503)
* Forward citation, reference lists & grey literature (n = 294)

**Identification**

Records screened

(n = 9003)

Records excluded\*\*

(n = 8384)

Unobtainable manuscripts

(n = 22)

Reports sought for retrieval

(n = 619)

**Screening**

Reports excluded: 345

* Not 1:1 treatment (n = 97)
* No effectiveness

data (n = 81)

* Not routine care (n = 69)
* No primary data (n = 64)
* Not psychotherapy (n = 14)
* Non adult population

(n = 5)

* Not face-to-face (n = 2)
* No English full text (n = 1)
* Other/NA (n = 6)

Reports assessed for eligibility

(n = 597)

Studies included in review

(n = 252)

Included in meta-analysis

(n = 223)

**Included**

| **Table 2:**  Summary statistics across the pooled sample and by sample severity. | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **university** | **mild** | **moderate** | **severe** | **other** | **total** | |
| N | N | 9195 | 158150 | 9515 | 22586 | 33694 | 233140 | |
| *k* | 58 | 88 | 32 | 92 | 8 | 278 | |
| mean | 158.53 | 1797.16 | 297.34 | 245.50 | 4211.75 | 838.63 | |
| median | 93.50 | 121.00 | 61.00 | 63.00 | 93.00 | 81.50 | |
| iqr | 162.50 | 935.00 | 347.00 | 107.50 | 1999.75 | 224.50 | |
| Females | N | 5350 | 95373 | 5797 | 14952 | 22801 | 144273 | |
| Age | *k* | 65.000 | 77.00 | 29 | 82 | 7 | 260 | |
| mean | 33.78 | 36.53 | 34.80 | 35.55 | 36.24 | 35.33 | |
| min | 20.50 | 19.00 | 24.30 | 21.52 | 24.52 | 19.00 | |
| max | 52.29 | 60.50 | 47.49 | 52.00 | 46.10 | 60.50 | |
| Sessions | *k* | 54 | 64 | 4 | 54 | 6 | 182 | |
| mean | 21.00 | 11.26 | 13.75 | 14.67 | 8.55 | 15.13 | |
| min | 2.15 | 4.00 | 9.00 | 1.00 | 8.00 | 1.00 | |
| max | 85.33 | 64.90 | 24.00 | 64.00 | 9.52 | 85.33 | |
| median | 14.77 | 8.18 | 11.00 | 11.15 | 8.15 | 13.00 | |
| iqr | 13.55 | 8.60 | 3.750 | 10.00 | 1.20 | 9.98 | |
| Setting | mixed | 0 | 0 | 0 | 0 | 5 | 5 | |
| outpatient | 68 | 96 | 0 | 91 | 4 | 259 | |
| inpatient | 0 | 0 | 33 | 1 | 0 | 34 | |
| Continent | Asia | 4 | 1 | 0 | 0 | 1 | 6 | |
| Australasia | 5 | 0 | 0 | 5 | 0 | 10 | |
| Europe | 20 | 13 | 15 | 14 | 1 | 63 | |
| America | 38 | 32 | 10 | 39 | 4 | 123 | |
| UK | 1 | 50 | 8 | 34 | 3 | 96 | |
| Analysis | inclusion | 48 | 48 | 16 | 53 | 4 | 169 | |
| completers | 19 | 45 | 16 | 35 | 3 | 118 | |
| therapy modality | cognitive-behavioural | 43 | 41 | 14 | 49 | 5 | 152 | |
| counselling | 0 | 22 | 0 | 3 | 0 | 25 | |
| psycho-dynamic | 12 | 9 | 13 | 16 | 0 | 50 | |
| other | 13 | 24 | 6 | 24 | 4 | 71 | |
| treatment stage | preliminary | 4 | 6 | 7 | 16 | 1 | 34 | |
| evaluations | 64 | 90 | 26 | 76 | 8 | 264 | |

### Sample Characteristics

Sample characteristics were reported for 291 samples, with a cumulative N of 233,140 (mean = 838.63, median = 81.5, range = 4 - 33,243, IQR = 224.5 ). The female preponderance was 61.88% (N = 144,273, *k* = 279) with 13 all female samples and 2 all male samples. The mean average sample age was 35.33 years (range = 19.00 - 60.50). Across studies which provided information, 23.00% of patients represented minority ethnic backgrounds (*k* = 127), 37.00% were married (*k* = 106), and 23.00% were in employment (*k* = 96).

### Treatment Characteristics

Most samples used a cognitive-behavioral treatment modality (*k* = 152) while 50 samples used psychodynamic, and 25 samples used counselling (other = 71). For severity, 96 (32.21%) samples came from mild services, 92 (30.87%) from moderate services, 33 (11.07%) from severe services, and 68 (22.82%) from University services (other, *k* = 9, 3.02%). Treatment dosage, when reported (*k* = 256) was in hours/sessions (*k* = 225), months (*k* = 12) or days (*k* = 8). The pooled (non-weighted) average dosage (hours) was 16.30 sessions (median = 13.00, range = 1.00-139.30, IQR = 11.00). 62 samples reported that treatment was delivered exclusively by unqualified/trainee clinicians, while 100 samples reported having at least one unqualified/trainee clinician.

### Risk of Bias

There was 60 studies rated across 8 criteria (i.e., 480 items). The pooled agreement was 84.17 ( = 0.62). Individual items varied in there levels of agreement (70.0%-98.3%). In order of most met criteria was demographic reporting detail (264/298, agreement = 98.33%, = 0.88), service reporting detail (260/298, agreement = 85%, = 0.31), study outcome reporting details (240/298, agreement = 83.33%, = -0.03), intervention reporting detail (234/298, agreement = 85%, = 0.32), service inclusion criteria (214/298, agreement = 90%, = 0.64), appropriate use of analysis (214/298, agreement = 70%, = 0.26), complete inclusion (i.e. consecutive recruitment and inclusion of those lost to follow-up, 41/298, agreement = 85%, = 0.45), and consecutive inclusion (93/298, agreement = 76.67%, = 0.51),

## Meta-Analyses

Details for individual studies (characteristics, effect-size, measurement tools) are presented in tabular form and forest plots (supplementary tables 4-5, supplementary figures 1-3). During the GRADE methodological appraisal process, each of the meta-analysis were initially rated as ‘low,’ based on the predominant type of study design within the available evidence. Following review of the five GRADE areas these overall ratings were reduced in level to ‘very low’ based on study limitations and also inconsistency within the available evidence.

The random-effects meta-analysis for depression outcomes (*k* = 140, N = 68,077), across 10 unique measurement tools was significant (*p* = < 0.001), indicative of a strong (*d* = 0.96, CI = 0.88-1.04) reduction in depression symptoms. There was significant study heterogeneity (I2 = 97.94%, Q[df = 121] = 2,677.37, *p* = < 0.001). The funnel plot (Figure 2) shows limited visual evidence of asymmetry. The funnel rank correlation test was not significant ( = 0.061, *p* = 0.46) however the funnel regression test was significant (Z = 2.13, *p* = 0.033). The fail-safe N was 515,853.

**Figure 2:**

Funnel plots displaying the distribution of studies reporting pre-post outcomes for (i) depression, (ii) anxiety, and (iii) miscellaneous outcomes.

**Chart, scatter chart

Description automatically generatedDepression**

**Chart, scatter chart

Description automatically generatedAnxiety**

**Chart, scatter chart

Description automatically generatedMiscellaneous**

Standardised Mean Change

The random-effects meta-analysis for anxiety outcomes (*k* = 84, N = 26,689, measurement tools = 20) was significant (*p* = < 0.001), indicative of a strong (*d* = 0.8, CI = 0.71-0.9) reduction in symptoms. There was significant heterogeneity across studies (I2 = 97.51%, Q[df = 68] = 1,328.96, *p* = < 0.001). The funnel plot shows limited evidence of asymmetry. The funnel rank correlation test was not significant ( = 0.009, *p* = 0.888). In contrast, the funnel regression test was significant (Z = 2.533, *p* = 0.011). The fail-safe N was 121,899

The random-effects meta-analysis for miscellaneous outcomes (i.e., global distress, functioning, diagnosis specific; *k* = 184, N = 126,734, measurement tools = 40) was significant (*p* = < 0.001), indicative of a strong (*d* = 1.01, CI = 0.93-1.09) reduction in symptoms. There was evidence of significant study heterogeneity across the included studies (I2 = 99.06%, Q[df = 157] = 15,330.32, *p* = < 0.001). The funnel plot shows a degree of asymmetry with clustering to the right of the mid-line. The funnel rank correlation test was significant ( = 0.208, *p* = <0.001). In contrast, the funnel regression test was not significant (Z = 3.697, *p* = <0.001). The fail-safe N was 1,695,607.

The post-hoc fixed effects model (i.e., sensitivity analysis) was highly comparable for miscellaneous outcomes (fixed: *d* = 1.02, CI = 1.01-1.02; random: *d* = 1.01, CI = 0.93-1.09). The discrepancy was larger for depression outcomes (fixed: *d* = 0.81, CI = 0.8-0.82; random: *d* = 0.96 CI = 0.88-1.04). and larger still for anxiety outcomes (fixed: *d* = 0.57, CI = 0.56-0.58; random: *d* = 0.8 CI 0.71-0.9).

## Moderator Analyses

Moderator analysis results are shown in Tables 3:6. The results of the omnibus models (QM tests) indicate moderators with significant between level variation. *Continent* was significant for all three outcome domains, with UK and North American studies producing larger effect sizes than other continents. Type of *analysis*, *sample size*, and treatment *develeopment stage* produced no significant findings. For *setting* (i.e., inpatient vs. outpatient) the omnibus model for anxiety was significant. Outpatient samples [*d* = 0.89, CI = 0.78-0.99] outperformed inpatient samples [*d* = 0.58, CI = 0.31-0.85]); although with CI overlap. Omnibus models for *modality*, *severity*, *experience* and *measurement tool* were significant for the anxiety and miscellaneous domains. Therapy *modality* demonstrated that counselling (*d* = 0.43, CI = 0.38-0.49) produced smaller improvements in anxiety, although with a small number of samples (*k* = 2). For miscellaneous outcomes CBT produced larger effect-sizes (*d* = 1.18, CI = 1.05-1.32) although with a narrow degree of overlap with psychodynamic (*d* = 0.93, CI = 0.79-1.07) and counselling outcomes (*d* = 0.90, CI = 0.75-1.06). For *severity*, anxiety outcomes were larger for mild (*d* = 0.99, CI = 0.79-1.20) and university services (*d* = 1.01, CI = 0.83-1.20) than for severe or moderate services. For miscellaneous outcomes, effect sizes were larger for mild (*d* = 1.08, CI = 0.95-1.21) and severe (*d* = 1.09, CI = 0.91-1.26) services. For *experience* unqualified professionals produced smaller improvements for miscellaneous outcomes (*d* = 0.77, CI = 0.65-0.89) with no overlap; although superior improvements for anxiety outcomes (*d* = 1.12, CI = 0.86-1.39). Finally, for *measurement tool*, larger effect-sizes for anxiety outcomes was shown for the GAD-7 (*d* = 0.96, CI = 0.78-1.15) than the BAI (*d* = 0.71, CI = 0.54-0.88). For miscellaneous outcomes, the OQ-45 demonstrated smaller effect-sizes (*d* = 0.57, CI = 0.41-0.74) than the other commonly used measures (BSI, CORE-OM, SCL).

**Table 3:**

Subgroup moderator analyses for depression outcomes.

|  |  | ***k*** | ***d*** | **ci** | **Q** | **I2** |
| --- | --- | --- | --- | --- | --- | --- |
| *Random effects model for severity (Q = 0.79, p = 0.853)* | | | | | | |
| severity | University | 30 | 0.98 | 1.16-0.79 | 43207.72 | 100.0% |
| mild | 34 | 1.03 | 1.22-0.85 | 9550225.84 | 100.0% |
| moderate | 57 | 0.98 | 1.11-0.86 | 80913.95 | 100.0% |
| severe | 15 | 0.91 | 1.12-0.7 | 237284.01 | 100.0% |
| *Random effects model for analysis (Q = 2.93, p = 0.087)* | | | | | | |
| analysis | include | 81 | 0.93 | 1.03-0.84 | 9730935.19 | 100.0% |
| Completers | 59 | 1.08 | 1.22-0.94 | 221034.47 | 100.0% |
| *Random effects model for setting (Q = 0.45, p = 0.503)* | | | | | | |
| setting | Outpatient | 121 | 0.99 | 1.08-0.91 | 9660158.70 | 100.0% |
| Inpatient | 16 | 0.92 | 1.12-0.72 | 240002.33 | 100.0% |
| ***Random effects model for continent (Q = 17.63, p = 0.001\*)*** | | | | | | |
| continent | North America | 58 | 1.00 | 1.11-0.88 | 640493.04 | 100.0% |
| UK | 44 | 1.10 | 1.26-0.94 | 3564904.55 | 100.0% |
| Europe | 29 | 0.95 | 1.12-0.77 | 60424.34 | 100.0% |
| Australasia | 4 | 0.67 | 1.01-0.33 | 7087.67 | 100.0% |
| Asia | 5 | 0.59 | 0.8-0.37 | 91.78 | 96.0% |
| *Random effects model for therapy modality (Q = 1.16, p = 0.763)* | | | | | | |
| therapy modality | Psychodynamic | 24 | 1.03 | 1.2-0.86 | 41984.43 | 100.0% |
| Counselling | 6 | 0.89 | 1.1-0.69 | 3471906.01 | 100.0% |
| cognitive-behavioural | 90 | 1.00 | 1.11-0.89 | 393181.50 | 100.0% |
| Other | 20 | 0.96 | 1.17-0.75 | 307948.08 | 100.0% |
| *Random effects model for treatment development stage (Q = 0.33, p = 0.566)* | | | | | | |
| treatment stage | routine evaluations | 118 | 1.00 | 1.09-0.91 | 9955258.72 | 100.0% |
| preliminary | 22 | 0.95 | 1.12-0.78 | 546.66 | 96.0% |
| *Random effects model for experience (Q = 1.41, p = 0.235)* | | | | | | |
| experience | Qualified | 121 | 1.01 | 1.1-0.92 | 9878905.77 | 100.0% |
| Trainees | 19 | 0.90 | 1.06-0.74 | 76281.83 | 100.0% |
| *Random effects model for measurement tool (Q = 0, p = 0.956)* | | | | | | |
| measure-ment tool | BDI | 34 | 1.02 | 1.18-0.86 | 121013.68 | 100.0% |
| PHQ-9 | 30 | 1.01 | 1.22-0.81 | 3495045.61 | 100.0% |
| *Random effects model for sample size (Q = 4.72, p = 0.094)* | | | | | | |
| sample size | large | 74 | 1.02 | 1.13-0.91 | 9949550.00 | 100.0% |
| small | 33 | 0.85 | 0.99-0.7 | 625.88 | 95.0% |
| medium | 33 | 1.08 | 1.26-0.89 | 5470.21 | 99.0% |
| *Note.* \* = < .05 | | | | | | |

**Table 4:**

Subgroup moderator analyses for anxiety outcomes.

|  |  | ***K*** | ***d*** | **ci** | **Q** | **I2** |
| --- | --- | --- | --- | --- | --- | --- |
| ***Random effects model for severity (Q = 17.15, p = 0.001\*)*** | | | | | | |
| severity | university | 29 | 1.01 | 1.2-0.83 | 32067.93 | 100.0% |
| mild | 22 | 0.99 | 1.2-0.79 | 334405.48 | 100.0% |
| moderate | 24 | 0.63 | 0.76-0.5 | 30702.22 | 100.0% |
| severe | 8 | 0.59 | 0.9-0.29 | 108223.63 | 100.0% |
| *Random effects model for analysis (Q = 1.82, p = 0.178)* | | | | | | |
| analysis | include | 58 | 0.81 | 0.93-0.69 | 517063.05 | 100.0% |
| Completers | 26 | 0.96 | 1.14-0.77 | 92492.19 | 100.0% |
| ***Random effects model for setting (Q = 4.34, p = 0.037\*)*** | | | | | | |
| setting | Outpatient | 75 | 0.89 | 0.99-0.78 | 440122.14 | 100.0% |
| inpatient | 9 | 0.58 | 0.85-0.31 | 157933.03 | 100.0% |
| ***Random effects model for continent (Q = 15.49, p = 0.004\*)*** | | | | | | |
| continent | North America | 32 | 0.91 | 1.1-0.72 | 230641.72 | 100.0% |
| UK | 25 | 0.89 | 1.09-0.7 | 115765.33 | 100.0% |
| Europe | 20 | 0.79 | 0.93-0.65 | 27960.45 | 100.0% |
| Australasia | 4 | 0.61 | 1.01-0.21 | 3781.78 | 100.0% |
| Asia | 3 | 0.59 | 0.68-0.49 | 3.33 | 40.0% |
| ***Random effects model for therapy modality (Q = 67.97, p = <0.001\*)*** | | | | | | |
| therapy modality | psychodynamic | 12 | 0.90 | 1.04-0.75 | 11879.61 | 100.0% |
| Counselling | 2 | 0.43 | 0.49-0.38 | 28.37 | 96.0% |
| cognitive-behavioural | 62 | 0.87 | 1-0.74 | 174811.70 | 100.0% |
| Other | 8 | 0.75 | 0.97-0.53 | 155956.84 | 100.0% |
| *Random effects model for treatment development stage (Q = 0.02, p = 0.885)* | | | | | | |
| treatment stage | routine evaluations | 74 | 0.85 | 0.96-0.74 | 635244.53 | 100.0% |
| preliminary studies | 10 | 0.87 | 1.14-0.61 | 639.27 | 99.0% |
| ***Random effects model for experience (Q = 5.77, p = 0.016\*)*** | | | | | | |
| experience | qualified | 66 | 0.78 | 0.88-0.68 | 502201.12 | 100.0% |
| trainees | 18 | 1.12 | 1.39-0.86 | 59913.29 | 100.0% |
| ***Random effects model for measurement tool (Q = 3.92, p = 0.048\*)*** | | | | | | |
| measure-ment tool | BAI | 19 | 0.71 | 0.88-0.54 | 36173.30 | 100.0% |
| GAD-7 | 19 | 0.96 | 1.15-0.78 | 241386.08 | 100.0% |
| *Random effects model for sample size (Q = 0.25, p = 0.885\*)* | | | | | | |
| sample size | large | 45 | 0.84 | 0.95-0.72 | 635426.98 | 100.0% |
| medium | 13 | 0.93 | 1.3-0.57 | 2069.64 | 99.0% |
| small | 26 | 0.84 | 1.03-0.66 | 898.19 | 97.0% |
| *Note.* \* = < .05 | | | | | | |

**Table 5:**

Subgroup moderator analyses for general outcomes.

|  |  | ***k*** | ***d*** | **ci** | **Q** | **I2** |
| --- | --- | --- | --- | --- | --- | --- |
| **Random effects model for severity (Q = 10.09, p = 0.018\*)** | | | | | | |
| severity | university | 28 | 0.82 | 0.95-0.7 | 44711.91 | 100.0% |
| mild | 61 | 1.08 | 1.21-0.95 | 39097618.60 | 100.0% |
| moderate | 62 | 0.98 | 1.12-0.84 | 111313.98 | 100.0% |
| severe | 27 | 1.09 | 1.26-0.91 | 66242.24 | 100.0% |
| Random effects model for analysis (Q = 1.8, p = 0.179) | | | | | | |
| analysis | include | 95 | 0.98 | 1.09-0.87 | 12421993.52 | 100.0% |
| Completers | 89 | 1.08 | 1.18-0.97 | 10540486.30 | 100.0% |
| Random effects model for setting (Q = 0.75, p = 0.386) | | | | | | |
| setting | Outpatient | 153 | 1.00 | 1.08-0.92 | 104250384.75 | 100.0% |
| inpatient | 28 | 1.08 | 1.25-0.91 | 67693.60 | 100.0% |
| **Random effects model for continent (Q = 13.53, p = 0.009\*)** | | | | | | |
| continent | UK | 68 | 1.02 | 1.13-0.92 | 92810921.73 | 100.0% |
| North America | 60 | 1.07 | 1.25-0.9 | 4606355.34 | 100.0% |
| Europe | 47 | 1.00 | 1.12-0.88 | 157855.31 | 100.0% |
| Australasia | 4 | 0.81 | 0.9-0.72 | 330.52 | 99.0% |
| Asia | 5 | 0.90 | 1.2-0.61 | 575.15 | 99.0% |
| **Random effects model for therapy modality (Q = 13.46, p = 0.004\*)** | | | | | | |
| therapy modality | cognitive-behavioural | 83 | 1.18 | 1.32-1.05 | 178432.53 | 100.0% |
| psychodynamic | 36 | 0.93 | 1.07-0.79 | 49292.89 | 100.0% |
| Counselling | 19 | 0.90 | 1.06-0.75 | 318722.49 | 100.0% |
| Other | 46 | 0.87 | 0.98-0.76 | 103700450.06 | 100.0% |
| Random effects model for treatment development stage (Q = 0.08, p = 0.781) | | | | | | |
| treatment stage | preliminary studies | 24 | 1.06 | 1.29-0.82 | 4742.50 | 100.0% |
| routine evaluations | 160 | 1.02 | 1.1-0.94 | 104380709.35 | 100.0% |
| **Random effects model for setting (Q = 0.75, p = < 0.001\*)** | | | | | | |
| experience | qualified | 158 | 1.07 | 1.16-0.99 | 104283036.81 | 100.0% |
| trainees | 26 | 0.77 | 0.89-0.65 | 71069.62 | 100.0% |
| **Random effects model for measurement tool (Q = 26.99, p = < 0.001\*)** | | | | | | |
| measure-ment tool | BSI-GSI | 26 | 0.87 | 1-0.73 | 9966.85 | 100.0% |
| CORE-OM | 35 | 1.04 | 1.18-0.9 | 72198841.45 | 100.0% |
| OQ-45 | 13 | 0.57 | 0.74-0.41 | 1090596.15 | 100.0% |
| SCL (Global) | 22 | 1.05 | 1.23-0.87 | 15779.11 | 100.0% |
| PCL | 12 | 1.29 | 1.61-0.97 | 3642.32 | 100.0% |
| Random effects model for sample size (Q = 1.03, p = 0.599\*) | | | | | | |
| sample size | large | 110 | 1.01 | 1.1-0.92 | 104379542.58 | 100.0% |
| medium | 38 | 1.11 | 1.3-0.92 | 4366.03 | 99.0% |
| small | 36 | 0.99 | 1.18-0.81 | 1001.45 | 97.0% |
| *Note.* \* = < .05 | | | | | | |

| **Table 6:** Meta-regression moderator variables (continuous) for depression, anxiety and miscellaneous outcome domains. | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Domain** | **Moderator** | **Mean (range)** | **k** | **B** | **CI** | **SE** | **Q** | *p* | **R2** |
| **depression** | Year (of publication) | 2012.16 (1988-2020) | 140 | 0.00 | 0.01–0.02 | 0.01 | 0.24 | 0.62 | 0.00 |
| mean age | 36.25 (19-60.5) | 126 | 0.00 | 0.01–0.01 | 0.01 | 0.11 | 0.74 | 0.00 |
| Sessions (mean) | 15.32 (1-46.5) | 84 | 0.01 | 0.02-0 | 0.01 | 1.54 | 0.22 | 0.47 |
| ethnicity (% minority) | 0.23 (0-0.66) | 62 | -0.09 | 0.5–0.69 | 0.30 | 0.09 | 0.76 | 0.00 |
| gender (% female) | 0.67 (0-1) | 132 | 0.00 | 0.4–0.39 | 0.20 | 0.00 | 0.99 | 0.00 |
| employment (% in full-time | 0.53 (0.05-1) | 46 | 0.31 | 0.97–0.35 | 0.34 | 0.85 | 0.36 | 0.00 |
| **anxiety** | Year (of publication) | 2012.93 (1999-2020) | 84 | 0.02 | 0.04-0 | 0.01 | 3.78 | 0.05 | 3.27 |
| mean age | 35.36 (19-60.5) | 79 | 0.00 | 0.01–0.01 | 0.01 | 0.01 | 0.91 | 0.00 |
| Sessions (mean) | 16 (1-46.5) | 52 | 0.00 | 0.01–0.01 | 0.01 | 0.03 | 0.85 | 0.00 |
| ethnicity (% minority) | 0.19 (0-0.59) | 40 | 0.31 | 1.62–0.99 | 0.66 | 0.22 | 0.64 | 0.00 |
| gender (% female) | 0.67 (0.078-1) | 79 | -0.27 | 0.31–0.85 | 0.30 | 0.82 | 0.36 | 0.00 |
| employment (% in full-time | 0.6 (0.05-1) | 29 | 0.95 | 1.55-0.35 | 0.31 | 9.64 | 0.00\* | 24.18 |
| **miscellaneous** | Year (of publication) | 2012.55 (2000-2020) | 184 | 0.00 | 0.02–0.01 | 0.01 | 0.07 | 0.79 | 0.00 |
| mean age | 35.15 (21.8-52.5) | 156 | 0.00 | 0.01–0.01 | 0.01 | 0.01 | 0.93 | 0.00 |
| Sessions (mean) | 15.5 (1-64.9) | 108 | -0.01 | 0–0.01 | 0.00 | 1.77 | 0.18 | 0.69 |
| ethnicity (% minority) | 0.26 (0-0.7) | 69 | -0.48 | 0.19–1.15 | 0.34 | 2.00 | 0.16 | 1.20 |
| gender (% female) | 0.67 (0-1) | 173 | -0.26 | 0.15–0.67 | 0.21 | 1.50 | 0.22 | 0.12 |
| employment (% in full-time | 0.53 (0-1) | 60 | -0.14 | 0.5–0.78 | 0.33 | 0.18 | 0.67 | 0.00 |
| *Note.*  \* p = <.05. | | | | | | | | | |

For the meta-regression (i.e., continuous) variables, there was no significant omnibus models for depression outcomes or miscellaneous outcomes. Employment was the only significant model for the anxiety domain.

For the analysis\*sessions multi-variate analysis, neither variable was significant when combined within a multi-variate model for any of the three outcome domains. These findings remained when allowing the variables to interact, with the interaction term also not significant. For loss to follow-up\*sessions, the multi-variate model found sessions to not be significant for each domain. Continent was not significant for the anxiety or miscellaneous domain. Continent was significant for the depression domain but only the level of the UK. When allowing for interaction terms there was no significant moderator or interaction term for any of the domains.

# Discussion

The aim of this review was to provide an up-to-date and systematic review of the literature for the effectiveness of psychological therapy in routine settings. 252 studies (*k* = 298) were identified, of which 223 (88.5%, *k* = 263) were eligible for the meta-analysis, producing the largest known synthesis of psychological therapy effectiveness studies conducted to date. The large number of included studies reflected the increased publication of practice-based evidence (71.48% since 2010). Consistent with prior psychotherapy effectiveness reviews, we found large pre-post effect sizes (*d* = 0.80 - 1.01) across multiple outcome domains (depression, anxiety, and miscellaneous outcomes). Continent was the only moderator significant for all domains, with UK and North American studies producing comparably larger effect-sizes.

For depression outcomes, no other moderator was significant in explaining heterogeneity. Four subgroup moderators (modality, severity, experience, measurement tool) were significant for anxiety and miscellaneous outcomes. For treatment modality, cognitive-behavioral interventions produced larger effect-sizes for miscellaneous outcomes. For treatment severity, anxiety effect-sizes were larger in mild and university services, while miscellaneous outcomes were larger in mild and severe services. For experience, training professionals’ effect-sizes were larger (i.e., compared to qualified professionals) for anxiety outcomes, but smaller for miscellaneous outcomes. Finally, for measurement tool, sensitivity to change was higher for the GAD-7 (anxiety) and lower for the OQ-45 (global distress). Anxiety outpatient samples also outperformed inpatient samples. The only significant meta-regressive variable was for employment (i.e., larger rates linked to larger effect-sizes) for anxiety samples. Multivariate meta-regressive models (analysis\*dosage, continent\*dosage) produced null findings.

## Interpretation of Findings

This review found that most individuals accessing psychological therapy are female, consistent with global epidemiological estimates of mental health gender prevalence (Seedat et al., 2009). The finding of large clinical improvements during psychotherapy and across outcomes was consistent with prior meta-analyses of psychotherapy effectiveness for depression outcomes ([Hans & Hiller, 2013](#ref-Hans2013); [Wakefield et al., 2021](#ref-Wakefield2021)), anxiety outcomes ([Stewart & Chambless, 2009](#ref-Stewart2009a); [Wakefield et al., 2021](#ref-Wakefield2021)), and nonspecific outcomes ([Cahill et al., 2010](#ref-Cahill2010)). Pooled effect-sizes were smaller than that found for [Cahill et al.](#ref-Cahill2010) ([2010](#ref-Cahill2010)) (*d* = 1.29), although this may reflect differences in review focus (e.g., Cahill et al., 2010 included group treatments) or changing distribution of geographical representation (i.e., more studies from non-UK/North American countries). Large clinical improvements are also in fitting with countless meta-analyses of psychotherapy controlled trials ([Cuijpers, Sijbrandij, et al., 2014](#ref-Cuijpers2014); e.g., [Cuijpers et al., 2008](#ref-Cuijpers2008); [Mayo-Wilson et al., 2014](#ref-Mayo-Wilson2014); [Olatunji et al., 2014](#ref-Olatunji2014)). The sensitivity analysis (i.e., fixed effects model) demonstrated a notable drop in anxiety outcomes (from *d* 0.80 to 0.57) although still represents important clinical improvements.

The finding that the UK and North America produce larger effect-sizes was a novel finding. It is possible that there are continental differences in models of training, service structures, therapy provision and emphasis on evidence-based practice which underlie the observed differences in pooled effect-sizes between continents. This is consistent with UK and US clinical guidance recommending delivery of empirically supported treatments ([APA, 2006](#ref-APA2006); [NICE, 2011](#ref-NICE2011)). Despite differences, all continents demonstrated positive change for all outcomes (*d* = 0.59 - 1.10) supporting the universality hypothesis (i.e., that psychotherapy is assumed to work across cultures., [Flückiger et al., 2018](#ref-Fluckiger2018)).

Consistent with prior evidence [Cuijpers et al.](#ref-Cuijpers2020) ([2020](#ref-Cuijpers2020)), psychotherapy produced strong clinical improvements for depression but with absence of evidence for significant between modality differences ([Barth et al., 2013](#ref-Barth2013a)). The finding that counselling produced smaller effect-sizes for anxiety outcomes was not expected and may be an artefact of few included samples (*k* = 2). Perhaps more surprising was that cognitive-behavioural approaches fared no better than psychodynamic approaches, despite CBT being a front-line treatment within various mental healthcare systems (e.g., [NICE, 2011](#ref-NICE2011)). Cognitive-behavioural approaches produced superior effect sizes for miscellaneous outcomes however it was not clear (without further analysis) if this was related to superiority for global distress outcomes, specific diagnosis outcomes (e.g., PTSD, OCD) or a combination as this domain was not specific.

Consistent with several prior meta-analytic reviews (e.g., [Cuijpers, Turner, et al., 2014](#ref-Cuijpers2014a); [Driessen et al., 2010](#ref-Driessen2010); [Furukawa et al., 2017](#ref-Furukawa2017)) sample severity did not predict effectiveness of treatment for depression. For anxiety and miscellaneous outcomes, services categorised ‘mild’consistently produced larger outcomes. For anxiety, university samples were also comparable, while for miscellaneous outcomes inpatient samples were comparable. In sum, effect-sizes for non-depression outcomes were consistently reduced for more severe community dwelling patients which provides support for increasing severity being associated with smaller improvements. What is less clear is the inconsistency for inpatient samples. For miscellaneous outcomes the inpatient setting produces comparable results (i.e., to outpatient) however for anxiety samples inpatient samples under performed. It is possible that classifying by type of service may have been an imprecise proxy for sample severity and therefore future research should explore severity as a continuous variable in routine settings.

An unexpected finding was that trainee clinicians produced significantly larger effect-sizes for anxiety and smaller effect-sizes for miscellaneous outcomes. Prior research has largely failed to find a significant difference between qualified and unquailed clinicians (e.g., [Boer et al., 2005](#ref-Boer2005); [Buckley et al., 2006](#ref-Buckley2006); [Montgomery et al., 2010](#ref-Montgomery2010)). A potential explanation is that clinicians in training are highly supervised and may therefore: (i) be less likely to ‘therapeutically drift’ ([Waller & Turner, 2016](#ref-Waller2016)) from the identified therapeutic model, (ii) receive more up-to-date training on evidence-based approaches; and (iii) more readily engage in deliberate skills practice. It is not clear however why this would apply to anxiety conditions but not miscellaneous outcomes. The current study failed to find support for dosage, sample size, gender, type of analysis, ethnicity, stage of study development, age, or publication as moderators of effect size.

## Limitations

Perhaps the most notable critique of this review is that it it based exclusively upon observational evidence (i.e., no control group or randomization). The absence of comparison conditions means that we are unable to rule out alternative explanations for observed effect-sizes such as placebo or natural recovery ([Posternak & Miller, 2001](#ref-Posternak2001); [Whiteford et al., 2012](#ref-Whiteford2012)).

A key design limitation surrounds statistical dependency. Efforts to avoid statistical dependency included: (i) taking one sample measure per domain, (ii) aggregating multiple unique study samples within a single domain, and (iii) extracting one measurement tool per study, per construct (i.e., preference system). These approaches have known limitations ([Borenstein et al., 2021](#ref-Borenstein2021); [Hoyt & Del Re, 2018](#ref-Hoyt2018); [Van den Noortgate et al., 2013](#ref-VandenNoortgate2013a)). A more sufficient approach would have been to model dependency using a multi-level analysis ([Van den Noortgate et al., 2013](#ref-VandenNoortgate2013a), [2015](#ref-VandenNoortgate2015)) and should be considered for replications.

An additional limitation is that bias was not sufficiently assessed. It became apparent during the review stage that the RoB tool had significant limitations; that it is primarily auditing manuscript reporting detail and not necessarily assessing risk of bias. In the absence of a sufficient assessment of RoB we could not determine the degree to which bias may have influenced results. Future reviews of effectiveness should strive to use the most appropriate RoB tool available (see [Munder & Barth, 2018](#ref-Munder2018)).

Due to resource constraints, the systematic search, data extraction and RoB ratings were not performed in duplicate. For the subsample of full texts screened by two coders there was a strong, but not perfect agreement/ reliability (80%, = 0.65). There is subsequently certainty that several studies screened only once will have been incorrectly excluded. Future reviews should follow best practice guidelines of screening in duplicate ([Polanin et al., 2019](#ref-Polanin2019)). Similarly, not extracting data or assessing RoB in duplicate is problematic due to risk of unreliable estimates of treatment effect and RoB ([Armijo-Olivo et al., 2014](#ref-Armijo-Olivo2014)). An additional limitation surrounds coding decisions for moderator variables. Therapy modality was coded from manuscript self-definition. The degree to which treatments truly resembled treatment code (or treatment intended) is not clear. It was also apparent during extraction that very few practice-based studies report fidelity/adherence checks. As this becomes more routinely reported opportunities for modelling differences based on fidelity/adherence will become available.

The search strategy used in this review, although produced many results, is unlikely to have identified every available study. Search terms were based on prior reviews and omitted several terms that were found to produce an unmanageable number of hits (e.g., “effectiveness,” “evaluation”). Despite this we feel that the current reviews gives an adequate range and depth of effectiveness research with which to make tentative interpretations regarding the field of psychotherapy effectiveness research.

A further caveat is the decision to focus exclusively on self-report measures of effectiveness. Meta-analytic evidence has demonstrated significant differences between self-report and clinician rated measures of clinical improvement ([Cuijpers et al., 2010](#ref-Cuijpers2010a)). Future research is therefore needed to see if the pooled effect-sizes from this study are consistent with clinician rated measures of effectiveness in routine settings.

## Implications for Research, Policy & Practice

To provide further understanding around the effectiveness of routinely delivered therapy future research should: (a) include fidelity and competency measures to confirm whether treatments delivered resembled treatment intended; (b) routinely assess outcomes at follow-up to establish maintenance of gains; (c) provide greater representation of therapy outcomes from non-western countries/services; and (d) explore variability in outcome between clinicians.

In terms of policy and practice, the following implications are considered. First, the need for development of reporting standards for practice-based evidence. The marked variation in how studies report details around the sample and intervention make comparisons and replication difficult. For example ethnicity rates were only reported for 127 samples (42.62%). This prevents accurate calculation of ethnicity rates across services/studies. Simply calculating the average rate of representation across those studies which do report statistics is not a valid approach as it is does not account for why studies omit ethnicity rates. Potential reasons include clinician/researcher oversight in reporting, or alternatively a marked lack of ethnic representation/access in these services/studies. There was also a lack of endeavor from studies to contextualize demographic utilization rates in terms of how representative they are of the populations/communities that they are intended to serve. Future practice-based studies of therapy effectiveness should routinely report all relevant rates of patient demographics and also quantify how proportionate they are of communities served.

Second, this study found no evidence of differential outcome based on ethnicity, age, or marital status through meta-regression. This provides further support for the need to provide fair and equitable access of psycholotherapy across the dimensions of age, ethnicity and marital status as there is no evidence that they impede effectiveness.

Third, routine recording of outcomes maintained at follow-up points should be enabled through necessary service commissioning of follow-up reviews/assessments. The body of evidence presented here concerns improvements made at the end of treatment. While follow-up was not the focus of this review, it was frequently apparent to reviewers that follow-up was rarely reported within studies. This information is necessary to determine the durability of improvements made during treatment.

Fourth, in light of differential outcomes demonstrated between qualified and unqualified clinicians (e.g. unqualified producing greater outcomes for anxiety) a review of training needs may be required for clinicians at different levels of experience.

## Conclusion

This review provides substantial support for the effectiveness of psychological therapy as delivered in routine settings across a range of outcomes. A key limitation of this review, and potentially the wider literature is the highly western-centric representation and reliance upon observational pre-post study designs. Nevertheless, for patients seeking help for psychological distress in routine services, there is growing evidence that interventions provided are clinically effective. The challenge for routine service delivery and associated effectiveness research is now to demonstrate the durability of this acute phase effect.

# References

Addis, M. E., & Krasnow, A. D. (2000). A national survey of practicing psychologists’ attitudes toward psychotherapy treatment manuals. *Journal of Consulting and Clinical Psychology*, *68*(2), 331–339. <https://doi.org/10.1037/0022-006X.68.2.331>

APA. (2006). Evidence-based practice in psychology. *Presidential Task Force on Evidence-Based Practice. American Psychologist*, *61*, 271–285. <https://doi.org/www.apa.org/pubs/journals/features/evidence-based-statement.pdf>

Armijo-Olivo, S., Ospina, M., Costa, B. R. da, Egger, M., Saltaji, H., Fuentes, J., Ha, C., & Cummings, G. G. (2014). Poor Reliability between Cochrane Reviewers and Blinded External Reviewers When Applying the Cochrane Risk of Bias Tool in Physical Therapy Trials. *PLOS ONE*, *9*(5), e96920. <https://doi.org/10.1371/journal.pone.0096920>

Balk, E. M., Earley, A., Patel, K., Trikalinos, T. A., & Dahabreh, I. J. (2012). Empirical assessment of within-arm correlation imputation in trials of continuous outcomes. *Methods Research Reports*, *12*(13).

Barkham, M., Margison, F., Leach, C., Lucock, M., Mellor-Clark, J., Evans, C., Benson, L., Connell, J., Audin, K., & McGrath, G. (2001). Service profiling and outcomes benchmarking using the CORE-OM: Toward practice-based evidence in the psychological therapies. *Journal of Consulting and Clinical Psychology*, *69*(2), 184–196. <https://doi.org/10.1037/0022-006X.69.2.184>

Barkham, M., Stiles, W. B., Lambert, M. J., & Mellor-Clark, J. (2010). Building a rigorous and relevant knowledge base for the psychological therapies. In M. Barkham, G. E. Hardy, & J. Mellor-Clark (Eds.), *Developing and Delivering Practice-Based Evidence* (pp. 21–61). John Wiley & Sons Ltd. <https://doi.org/10.1002/9780470687994.ch2>

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., Jüni, P., & Cuijpers, P. (2013). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: A network meta-analysis. *PLoS Medicine*, *10*(5), e1001454. <https://doi.org/10.1371/journal.pmed.1001454>

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical & Statistical Psychology*, *41*(2), 257–278. <https://doi.org/10.1111/j.2044-8317.1988.tb00901>

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*(4), 1088–1101. <https://doi.org/10.2307/2533446>

Boer, P., Wiersma, D., Russo, S., & Bosch, R. P. (2005). Cognitive self-therapy: A contribution to long-term treatment of depression and anxiety. *Cochrane Database for Systematic Reviews*, *2*.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2021). *Introduction to Meta-Analysis*. John Wiley & Sons.

Buckley, J. V., Newman, D. W., Kellett, S., & Beail, N. (2006). A naturalistic comparison of the effectiveness of trainee and qualified clinical psychologists. *Psychology and Psychotherapy: Theory, Research and Practice*, *79*(1), 137–144. <https://doi.org/10.1348/147608305X52595>

Cahill, J., Barkham, M., & Stiles, W. (2010). Systematic review of practice-based research on psychological therapies in routine clinic settings. *The British Journal of Clinical Psychology*, *49*(4), 421–453. <https://doi.org/10.1348/014466509X470789>

Card, N. A. (2015). *Applied Meta-Analysis for Social Science Research*. Guilford Publications.

Carpenter, J. K., Andrews, L. A., Witcraft, S. M., Powers, M. B., Smits, J. A. J., & Hofmann, S. G. (2018). Cognitive behavioral therapy for anxiety and related disorders: A meta-analysis of randomized placebo-controlled trials. *Depression and Anxiety*, *35*(6), 502–514. <https://doi.org/10.1002/da.22728>

Castonguay, L. G., Barkham, M., Jeong Youn, S., & Page, A. C. (2021). Practice-based evidence findings from routine clinical settings. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield’s Handbook of Psychotherapy and Behavior Change* (Seventh, pp. 191–222). Wiley.

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavey, A. A. (2013). Practice-oriented reserach: Approaches and applications. In M. J. Lambert (Ed.), *Bergin and Garfield’s Handbook of Psychotherapy and Behavior Change* (Sixth, pp. 85–133). Wiley.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology,* *66*(1), 7–18. <https://doi.org/10.1037//0022-006x.66.1.7>

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Empirically Supported Psychological Interventions: Controversies and Evidence*, *52*(1), 685–716. <https://doi.org/10.1146/annurev.psych.52.1.685>

Clark, D. M., Canvin, L., Green, J., Layard, R., Pilling, S., & Janecka, M. (2018). Transparency about the outcomes of mental health services (IAPT approach): An analysis of public data. *The Lancet*, *391*(10121), 679–686. <https://doi.org/10.1016/S0140-6736(17)32133-5>

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*(1), 101–129. <https://doi.org/10.2307/3001666>

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. <https://doi.org/10.1177/001316446002000104>

Connell, J., Barkham, M., & Mellor-Clark, J. (2007). CORE-OM mental health norms of students attending university counselling services benchmarked against an age-matched primary care sample. *British Journal of Guidance & Counselling*, *35*(1), 41–57. <https://doi.org/10.1080/03069880601106781>

Cooper, H. M. (1998). *Synthesizing Research: A Guide for Literature Reviews*. SAGE.

Cuijpers, P. (2016). *Meta-analyses in mental health research: A practical guide*. University of Amsterdam.

Cuijpers, P., Clignet, F., van Meijel, B., van Straten, A., Li, J., & Andersson, G. (2011). Psychological treatment of depression in inpatients: A systematic review and meta-analysis. *Clinical Psychology Review*, *31*(3), 353–360. <https://doi.org/10.1016/j.cpr.2011.01.002>

Cuijpers, P., Huibers, M., Daniel Ebert, D., Koole, S. L., & Andersson, G. (2013). How much psychotherapy is needed to treat depression? A metaregression analysis. *Journal of Affective Disorders*, *149*(1-3), 1–13. <https://doi.org/10.1016/j.jad.2013.02.030>

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

Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review*, *30*(6), 768–778. <https://doi.org/10.1016/j.cpr.2010.06.001>

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., Straten, A. van, 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., Turner, E. H., Mohr, D. C., Hofmann, S. G., Andersson, G., Berking, M., & Coyne, J. (2014). Comparison of psychotherapies for adult depression to pill placebo control groups: A meta-analysis. *Psychological Medicine*, *44*(4), 685–695. <https://doi.org/10.1017/S0033291713000457>

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>

de Jong, K., Conijn, J. M., Gallagher, R. A. V., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, *85*, 102002. <https://doi.org/10.1016/j.cpr.2021.102002>

Delgadillo, J., McMillan, D., Leach, C., Lucock, M., Gilbody, S., & Wood, N. (2014). Benchmarking routine psychological services: A discussion of challenges and methods. *Behavioural and Cognitive Psychotherapy*, *42*(1), 16–30. <https://doi.org/10.1017/S135246581200080X>

Delgadillo, J., McMillan, D., Leach, C., Lucock, M., Gilbody, S., & Wood, N. (2014). Benchmarking routine psychological services: A discussion of challenges and methods. *Behavioural and Cognitive Psychotherapy*, *42*(1), 16–30. <https://doi.org/10.1017/S135246581200080X>

Driessen, E., Cuijpers, P., Hollon, S. D., & Dekker, J. J. M. (2010). Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-analysis. *Journal of Consulting and Clinical Psychology*, *78*(5), 668–680. <https://doi.org/10.1037/a0020570>

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>

Flückiger, C., Del Re, A. C., Barth, J., Hoyt, W. T., Levitt, H., Munder, T., Spielmans, G. I., Swift, J. K., Vîslă, A., & Wampold, B. E. (2018). Considerations of how to conduct meta-analyses in psychological interventions. *Psychotherapy Research*, *28*(3), 329–332. <https://doi.org/10.1080/10503307.2018.1430390>

Flückiger, C., Wampold, B. E., Delgadillo, J., Rubel, J., Vîslă, A., & Lutz, W. (2020). Is there an evidence-based number of sessions in outpatient psychotherapy? A comparison of naturalistic conditions across countries. *Psychotherapy and Psychosomatics*, *89*(5), 333–335. <https://doi.org/10.1159/000507793>

Freedland, K. E., Mohr, D. C., Davidson, K. W., & Schwartz, J. E. (2011). Usual and unusual care: Existing practice control groups in randomized controlled trials of behavioral interventions. *Psychosomatic Medicine*, *73*(4), 323–335. <https://doi.org/10.1097/PSY.0b013e318218e1fb>

Furukawa, T. A., Weitz, E. S., Tanaka, S., Hollon, S. D., Hofmann, S. G., Andersson, G., Twisk, J., DeRubeis, R. J., Dimidjian, S., Hegerl, U., Mergl, R., Jarrett, R. B., Vittengl, J. R., Watanabe, N., & Cuijpers, P. (2017). Initial severity of depression and efficacy of cognitivebehavioural therapy: Individual-participant data meta-analysis of pill-placebo-controlled trials. *British Journal of Psychiatry*, *210*(3), 190–196. <https://doi.org/10.1192/bjp.bp.116.187773>

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (Second, pp. 357–376). Russell Sage.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, *336*(7650), 924–926. <https://doi.org/10.1136/bmj.39489.470347.AD>

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

Hans, E., & Hiller, W. (2013). Effectiveness of and dropout from outpatient cognitive behavioral therapy for adult unipolar depression: A meta-analysis of nonrandomized effectiveness studies. *Journal of Consulting and Clinical Psychology*, *81*(1), 75–88. <https://doi.org/10.1037/a0031080>

Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, David. D. (2019a). *Dmetar: Companion R package for the guide ’doing meta-analysis in R’.* [R Package].

Harrer, M., Cuijpers, P., Furukawa, Toshi. A., & Ebert, David. D. (2019b). Multiple Meta-Regression. In *Doing meta-analysis in R: A hands-on guide*.

Higgins, J., & Green, S. (Eds.). (2008). *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9780470712184.fmatter>

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

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

Hoyt, W. T., & Del Re, A. C. (2018). Effect size calculation in meta-analyses of psychotherapy outcome research. *Psychotherapy Research*, *28*(3), 379–388. <https://doi.org/10.1080/10503307.2017.1405171>

Hunsley, J., & Lee, C. M. (2007). Research-informed benchmarks for psychological treatments: Efficacy studies, effectiveness studies, and beyond. *Professional Psychology: Research and Practice*, *38*(1), 21–33. <https://doi.org/10.1037/0735-7028.38.1.21>

Jüni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of controlled clinical trials. *BMJ*, *323*(7303), 42–46.

Lambert, M. J. (2013). The efficacy and effectiveness of psychotherapy. In M. J. Lambert & A. E. Bergin (Eds.), *Bergin and Garfield’s handbook of psychotherapy and behavior change* (6th ed., pp. 169–218). John Wiley & Sons, Incorporated.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. <https://doi.org/10.2307/2529310>

Lemma, A., Roth, A. D., & Pilling, S. (2008). *The competences required to deliver effective Psychoanalytic/ Psychodynamic Therapy*. Research Department of Clinical, Educational and Health Psychology, University College London.

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), 1729633. <https://doi.org/10.1080/20008198.2020.1729633>

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), 1729633. <https://doi.org/10.1080/20008198.2020.1729633>

Linardon, J., Wade, T. D., de la Piedad Garcia, X., & Brennan, L. (2017). The efficacy of cognitive-behavioral therapy for eating disorders: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *85*(11), 1080–1094. <https://doi.org/10.1037/ccp0000245>

Lutz, W., Schiefele, A.-K., Wucherpfennig, F., Rubel, J., & Stulz, N. (2016). Clinical effectiveness of cognitive behavioral therapy for depression in routine care: A propensity score based comparison between randomized controlled trials and clinical practice. *Journal of Affective Disorders*, *189*(1), 150–158. <https://doi.org/10.1016/j.jad.2015.08.072>

Margison, F. R., Barkham, M., Evans, C., McGrath, G., Clark, J. M., Audin, K., & Connell, J. (2000). Measurement and psychotherapy: Evidence-based practice and practice-based evidence. *British Journal of Psychiatry*, *177*(2), 123–130. <https://doi.org/10.1192/bjp.177.2.123>

Mayo-Wilson, E., Dias, S., Mavranezouli, I., Kew, K., Clark, D. M., Ades, A. E., & Pilling, S. (2014). Psychological and pharmacological interventions for social anxiety disorder in adults: A systematic review and network meta-analysis. *The Lancet Psychiatry*, *1*(5), 368–376. <https://doi.org/10.1016/S2215-0366(14)70329-3>

Montgomery, E. C., Kunik, M. E., Wilson, N., Stanley, M. A., & Weiss, B. (2010). Can paraprofessionals deliver cognitive-behavioral therapy to treat anxiety and depressive symptoms? *Bulletin of the Menninger Clinic*, *74*(1), 45–62. <https://doi.org/10.1521/bumc.2010.74.1.45>

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, *11*(2), 364–386. <https://doi.org/10.1177/1094428106291059>

Munder, T., & Barth, J. (2018). Cochrane’s risk of bias tool in the context of psychotherapy outcome research. *Psychotherapy Research*, *28*(3), 347–355. <https://doi.org/10.1080/10503307.2017.1411628>

Munn, Z., Barker, T. H., Moola, S., Tufanaru, C., Stern, C., McArthur, A., Stephenson, M., & Aromataris, E. (2020). Methodological quality of case series studies: An introduction to the JBI critical appraisal tool. *JBI Evidence Synthesis*, *18*(10), 2127–2133. <https://doi.org/10.11124/JBISRIR-D-19-00099>

National Collaborating Centre for Mental Health. (2011). *Common Mental Health Disorders: Identification and Pathways to Care* (Vol. 123). RCPsych Publications.

NICE. (2011). *Common mental health problems: Identification and pathways to care*. National Institute for Clinical Excellence (NICE).

Nordmo, M., Sønderland, N. M., Havik, O. E., Eilertsen, D.-E., Monsen, J. T., & Solbakken, O. A. (2020). Effectiveness of open-ended psychotherapy under clinically representative conditions. *Frontiers in Psychiatry*, *11*, 384. <https://doi.org/10.3389/fpsyt.2020.00384>

Olatunji, B. O., Davis, M. L., Powers, M. B., & Smits, J. A. J. (2015). Cognitive-behavioral therapy for obsessive-compulsive disorder: A meta-analysis of treatment outcome and moderators. *FOCUS*, *13*(2), 222–231. <https://doi.org/10.1176/appi.focus.130219>

Olatunji, B. O., Kauffman, B. Y., Meltzer, S., Davis, M. L., Smits, J. A. J., & Powers, M. B. (2014). Cognitive-behavioral therapy for hypochondriasis/health anxiety: A meta-analysis of treatment outcome and moderators. *Behaviour Research and Therapy*, *58*, 65–74. <https://doi.org/10.1016/j.brat.2014.05.002>

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan: A web and mobile app for systematic reviews. *Systematic Reviews*, *5*(1), 210. <https://doi.org/10.1186/s13643-016-0384-4>

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*, *18*(3), e1003583. <https://doi.org/10.1371/journal.pmed.1003583>

Persons, J. B., Bostrom, A., & Bertagnolli, A. (1999). Results of randomized controlled trials of cognitive therapy for depression generalize to private practice. *Cognitive Therapy and Research*, *23*(5), 535–548. <https://doi.org/10.1023/A:1018724505659>

Philips, B., & Falkenström, F. (2021). What research evidence is valid for psychotherapy research? *Frontiers in Psychiatry*, *11*, 1549. <https://doi.org/10.3389/fpsyt.2020.625380>

Polanin, J. R., Pigott, T. D., Espelage, D. L., & Grotpeter, J. K. (2019). Best practice guidelines for abstract screening large-evidence systematic reviews and meta-analyses. *Research Synthesis Methods*, *10*(3), 330–342. <https://doi.org/10.1002/jrsm.1354>

Posternak, M. A., & Miller, I. (2001). Untreated short-term course of major depression: A meta-analysis of outcomes from studies using wait-list control groups. *Journal of Affective Disorders*, *66*(2-3), 139–146. <https://doi.org/10.1016/s0165-0327(00)00304-9>

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Rosa-Alcázar, A. I., Sánchez-Meca, J., Gómez-Conesa, A., & Marín-Martínez, F. (2008). Psychological treatment of obsessivecompulsive disorder: A meta-analysis. *Clinical Psychology Review*, *28*(8), 1310–1325. <https://doi.org/10.1016/j.cpr.2008.07.001>

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

Roth, A., Hill, A., & Pilling, S. (2009). *The competences required to deliver effective Humanistic Psychological Therapies*. London: University College London.

Roth, A., & Pilling, S. (2008). Using an evidence-based methodology to identify the competences required to deliver effective cognitive and behavioural therapy for depression and anxiety disorders. *Behavioural and Cognitive Psychotherapy*, *36*(2), 129–147. <https://doi.org/10.1017/S1352465808004141>

Sánchez-Meca, J., Rosa-Alcázar, A. I., Marín-Martínez, F., & Gómez-Conesa, A. (2010). Psychological treatment of panic disorder with or without agoraphobia: A meta-analysis. *Clinical Psychology Review*, *30*(1), 37–50. <https://doi.org/10.1016/j.cpr.2009.08.011>

Schwarzer, G. (2020). *Meta: General package for meta-analysis*. <https://CRAN.R-project.org/package=meta>

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crits-Christoph, P., Hazelrigg, M. D., Jorm, A. F., Lyons, L. C., Nietzel, M. T., Robinson, L., Prout, H. T., Smith, M. L., Svartberg, M., & Weiss, B. (1997). Evidence that therapy works in clincally representative conditions. *Journal of Consulting and Clinical Psychology*, *65*(3), 355–365. <https://doi.org/10.1037/0022-006X.65.3.355>

Shadish, W. R., Navarro, A. M., Matt, G. E., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*(4), 512–529. <https://doi.org/10.1037/0033-2909.126.4.512>

Spielmans, G. I., & Flückiger, C. (2018). Moderators in psychotherapy meta-analysis. *Psychotherapy Research*, *28*(3), 333–346. <https://doi.org/10.1080/10503307.2017.1422214>

Stewart, R. E., & Chambless, D. L. (2009). Cognitive-behavioral therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies. *Journal of Consulting and Clinical Psychology*, *77*(4), 595–606. <https://doi.org/10.1037/a0016032>

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>

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2013). Three-level meta-analysis of dependent effect sizes. *Behavior Research Methods*, *45*(2), 576–594. <https://doi.org/10.3758/s13428-012-0261-6>

Van den Noortgate, W., López-López, J. A., Marín-Martínez, F., & Sánchez-Meca, J. (2015). Meta-analysis of multiple outcomes: A multilevel approach. *Behavior Research Methods*, *47*(4), 1274–1294. <https://doi.org/10.3758/s13428-014-0527-2>

van der Lem, R., de Wever, W. W., van der Wee, N. J., van Veen, T., Cuijpers, P., & Zitman, F. G. (2012). The generalizability of psychotherapy efficacy trials in major depressive disorder: An analysis of the influence of patient selection in efficacy trials on symptom outcome in daily practice. *BMC Psychiatry*, *12*(1), 192. <https://doi.org/10.1186/1471-244X-12-192>

Viechtbauer, W. (2020). *Metafor: Meta-analysis package for r*. <https://CRAN.R-project.org/package=metafor>

Wakefield, S., Kellett, S., Simmonds-Buckley, M., Stockton, D., Bradbury, A., & Delgadillo, J. (2021). Improving Access to Psychological Therapies (IAPT) in the United Kingdom: A systematic review and meta-analysis of 10-years of practice-based evidence. *British Journal of Clinical Psychology*, *60*(1), 1–37. <https://doi.org/10.1111/bjc.12259>

Waller, G., & Turner, H. (2016). Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy*, *77*, 129–137. <https://doi.org/10.1016/j.brat.2015.12.005>

Weisz, J. R., Weiss, B., Han, S. S., Granger, D. A., & Morton, T. (1995). Effects of psychotherapy with children and adolescents revisited: A meta-analysis of treatment outcome studies. *Psychological Bulletin*, *117*(3), 450–468. <https://doi.org/10.1037/0033-2909.117.3.450>

Whiston, A., Bockting, C. L. H., & Semkovska, M. (2019). Towards personalising treatment: A systematic review and meta-analysis of face-to-face efficacy moderators of cognitive-behavioral therapy and interpersonal psychotherapy for major depressive disorder. *Psychological Medicine*, *49*(16), 2657–2668. <https://doi.org/10.1017/S0033291719002812>

Whiteford, H., Harris, M., Mckeon, G., Baxter, A., Pennell, C., Barendregt, J., & Wang, J. (2012). Estimating remission from untreated major depression: A systematic review and meta-analysis. *Psychological Medicine*, *43*, 1–17. <https://doi.org/10.1017/S0033291712001717>

Wolitzky-Taylor, K. B., Horowitz, J. D., Powers, M. B., & Telch, M. J. (2008). Psychological approaches in the treatment of specific phobias: A meta-analysis. *Clinical Psychology Review*, *28*(6), 1021–1037. <https://doi.org/10.1016/j.cpr.2008.02.007>

Zimmerman, M., Mattia, J. I., & Posternak, M. A. (2002). Are subjects in pharmacological treatment trials of depression representative of patients in routine clinical practice? *American Journal of Psychiatry*, *159*(3), 469–473. <https://doi.org/10.1176/appi.ajp.159.3.469>

1. Protocol available at: [↑](#footnote-ref-1)