Contents

University of Sheffield

The Effectiveness of Psychotherapy Delivered in Routine Service Settings

Chris Gaskell

Supervisors:

Dr. Stephen C. Kellett

Dr. Mel Simmonds-Buckley

Dr. Jaime Delgadillo

A report submitted in partial fulfillment of the requirements
for the degree of Doctorate in Clinical Psychology
in the Department of Clinical Psychology

May 14, 2021

This page is intentionally left blank

Preface

Declaration

I declare that this work has not been submitted for any other degree at the University of Sheffield, or any other institution. The work presented is original and all other sources have been referenced accordingly.

This page is intentionally left blank

About

This is the Preliminary Manuscript for my meta-analysis on naturalistic psychotherapy outcomes, submitted in partial fulfillment for the DClinPsy qualification. This manuscript has been produced using R (R Core Team, 2020) with a combination of R Markdown (Allaire et al., 2020) and LaTeX. In addition to providing high quality dynamic reports R Markdown offers the advantage of maintaining a fully reproducible workflow. R Bookdown (Allaire et al., 2020) was used to compile the document; an accompanying dynamic html book is also available (`https://bookdown.org/cgaskell1/bookdown-demo/`).

This page is intentionally left blank

Word Counts

A crude word count for the combined in-text sections drafted so far is 8288. This word count was calculated using the R package Wordcountaddin which includes tables/figures. A more precise word count will be calculated for the final draft.

This page is intentionally left blank

Acknowledgements

This page is intentionally left blank

Contents

PART I

SYSTEMATIC LITERATURE REVIEW

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

Chris Gaskell

Supervisors:

Dr. Stephen C. Kellett

Dr. Mel Simmonds-Buckley

Abstract

Objectives: There has been a substantial increase in the amount of evidence arising from routine (i.e. naturalistic) care settings in the field of psychotherapy in recent decades. This review sought to examine the effectiveness of routinely delivered psychological therapies. Design: A pre-registered systematic-review and meta-analysis (CRD42020175235). Methods: Random-effects meta-analyses were conducted on studies meeting pre-specified inclusion criteria. Moderator analyses examining methodological, treatment-level and sample-level variables explored between-study heterogeneity. Results: The systematic search identified 252 studies (k = 298 samples) for the quantitative synthesis. Of these, 223 studies (k = 263 samples) were eligible for inclusion in a meta-analysis of pre-post treatment outcomes. Results showed large pre-post treatment effects for depression (d = 0.98, [CI 0.9-1.06], p = < 0.001, k = 140), anxiety (d = 0.83 [CI 0.73-0.92], p = < 0.001, k = 84), and global outcome domains (d = 1.01 [CI 0.93-1.08], p = < 0.001, k = 184). Sample completion (completers vs. intention-to-treat), geographical area (continent) and methodological quality were significant moderators of treatment effects. Conclusions: This review provides further 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 Outcome Monitoring,' 'Meta-analysis.'

Practitioner Points:

- Greater depth and consistency of reporting detail is required in routine outcome studies.
- Completer analyses may artificially inflate effect-sizes for outcomes in routine practice.
- There was encouraging evidence that age and ethnicity does not hinder treatment effectiveness; equitable opportunity to access treatment should therefore be provided across these dimensions.

1

Effectiveness of Tertiary Care Outpatient Psychological Interventions; a benchmarking study

There is widespread consensus that psychological therapy is an efficacious treatment for a variety of mental health disorders (Lambert, 2013). A substantial proportion of the evidence for these claims originates in reviews of randomised controlled trials (RCTs, e.g., Smith & Glass, 1977). The primary critique of these RCTs is that methods used to enhance experimental control (e.g. homogeneous client groups, random assignment, control groups) mean that the results may not necessarily generalize to routine services that typically treat a heterogeneous patient population (Barkham, Stiles, et al., 2010). Until recently there has been an over-reliance on this form of evidence (i.e. efficacy evidence), exemplified through a seminal review of studies (Roth & Fonagy, 1996) being criticized for being almost exclusively made up of RCTs (Margison et al., 2000). The extent to which efficacious treatments hold up in routine settings (i.e. transportability) remains a contentious debate in psychology (Hunsley & Lee, 2007; Jacobson & Christensen, 1996; Smith & Glass, 1977).

Routine service settings differ substantially to the conditions typically provided for efficacy research (Barkham, Stiles, et al., 2010). Routine services traditionally have higher patient to clinician ratios, higher levels of patient heterogeneity and manage greater levels of clinical risk. Interventions provided within these settings are less standardised, with less frequent use of protocols/manuals, scarce use of integrity/fidelity checks, and often the setting of limits (in terms of number of sessions possible).

The nature of interventions, how they are delivered, and also how they are evaluated within routine settings has changed over time. Evidence of psychological therapy outcome, arising from routine service is 'effectiveness' research, also known as 'practice-based evidence' (PBE: Barkham, Hardy, et al., 2010). A common critique

of PBE is the uncontrolled presence of potentially confounding variables, which may compromise internal validity (Barkham, Stiles, et al., 2010). Since the initial binary distinction of efficacy vs. effectiveness, it has become increasingly recognised that the two approaches can overlap (Stewart & Chambless, 2009), forming an efficacy-effectiveness continuum (Hunsley & Lee, 2007).

Models have been provided to consider how effectiveness and efficacy research may complement each other on this continuum. One example of this is the three-stage 'hour-glass' model of psychological therapy outcomes research (Salkovskis, 1995). Stage one consists of emerging intervention evidence conducted on small numbers of patients in routine practice, often using uncontrolled research designs (e.g. pilot and case studies). Stage two elaborates on promising stage one evidence by further investigating the intervention under more tightly controlled efficacy conditions (ideally RCT). Finally stage three involves the transporting of interventions, empirically supported at stage two, to larger practice-based (naturalistic) settings in order to confirm/refute clinical utility. This is the evidence-based practice phase of the hourglass. This cycle is then repeated as the intervention is iterated and applied to new settings/populations.

Given that the majority of therapy is delivered within routine practice settings it is subsequently necessary to conduct regular reviews of the emerging evidence base generated within these settings (i.e. evidence from stages 1 and 3 of the hour-glass model). Prior reviews have employed meta-analytic approaches in order to aggregate effect-sizes across studies (see Lambert, 2013 for a review). Reviews of effectiveness research generally employ one of two approaches. The representativeness approach (e.g. Shadish et al., 2000; Smith & Glass, 1977; Stewart & Chambless, 2009) uses broad inclusion criteria – including efficacy studies – before rating each study on how much the conditions reported resemble routine services. Representativeness is then assessed for the degree to which it is associated with outcome. An alternative strategy is to restrict inclusion to studies which are highly representative of routine conditions (e.g. Cahill et al., 2010; Wakefield et al., 2021). For example Cahill et al. (2010) in their

review of 31 studies found support for the effectiveness of therapy conducted in routine settings, with more methodologically rigorous studies producing greater effect-sizes. In the 11 years since the last broad review of therapy effectiveness (Cahill et al., 2010) there has been a considerable increase in the volume of practice-based evidence, thus justifying the need for an updated review. Furthermore, although earlier reviews have quantitatively examined the effectiveness of routinely-delivered therapy, there is scarce evidence on sources of heterogeneity of treatment effects.

Heterogeneity refers to the amount of variability inherent in the aggregated treatment effect-size; and subsequently influences the degree to which findings can be confidently generalized (Kraemer et al., 2006). Patient heterogeneity is generally higher in practice-based studies (compared with RCTs) because of the less frequent use of exclusion criteria. Heterogeneity within meta-analyses of effectiveness research is typically high (e.g. Wakefield et al., 2021). A treatment effect with high heterogeneity may fail to explain potential underlying differences in how different people respond to treatment.

A common approach is to try to reduce the unexplained heterogeneity by measuring moderator variables that may account for a proportion of the heterogeneity. A moderator is a pre-treatment variable that can be used to define subgroups of patients within a larger sample (Kraemer et al., 2006). A moderator of treatment effect is when differential rates of effectiveness between individuals is demonstrated based on prior distinction. Use of moderator variables has been somewhat limited in prior reviews of PBE. For example, Wakefield et al. (2021) reviewed studies conducted in the UK increasing access to psychological therapy (IAPT) programme and found that study methodology (intention-to-treat vs. completers analysis) was a significant and consistent moderator of treatment outcomes. The current review sought to measure the influence of a range pre-identified moderator variables. A hypothesis-led approach, based on the extant literature of psychological therapy outcomes was used to select moderator variables.

Moderators can be conceptualised as being at the levels of (i) patient, (ii) treatment, (iii) service or (iv) study methodology. Patient level moderators may include demographics (age, gender, ethnicity etc.) or presenting problem/diagnosis (e.g. Roth & Fonagy, 1996). Treatment level moderators may include therapeutic model (e.g. Roth & Fonagy, 1996), treatment dosage (e.g. Flückiger et al., 2020) or interventionist experience (e.g. Buckley et al., 2006). Service level moderators may include type of service, setting (inpatient vs. outpatient), geographical region, sector (e.g. Wakefield et al., 2021), or service funding structure. Finally, methodological variables may include year of publication, sample analysed (e.g. intention-to-treat, Wakefield et al., 2021), stage of the hour-glass or study methodological quality (e.g. Wakefield et al., 2021). These listed moderators were considered for inclusion in the current review.

When exploring multiple moderator variables it is important to consider the potential interactive relationships that they have. This is possible through a process called multi-variate moderator analysis. Considering only each moderator in isolation will not show how each moderator can become amplified or attenuated when considered with another (Li et al., 2020). The primary barrier to multi-variate moderator analysis is that these methods require a high ratio of studies to co-variates (Borenstein et al., 2009) which are often not available to researchers. A broad review of effectiveness studies is now required which can allow for multi-variate analysis of moderators of treatment effect-size.

Aims

The main objective of the present study was to qualitatively and quantitatively synthesize the evidence on the effectiveness of individual psychological therapy for adults accessing services in routine care. The primary aim was to quantify the effectiveness of psychological treatment delivered in routine services. In doing so, the present study used a liberal conceptualisation of what constitutes as a 'routine service' which matches the reality of the inherent heterogeneity shown across routine care settings. This included a systematic search and meta-analysis of studies published prior

to the systematic search date. The secondary aim of this review was to explore how a range of moderator variables influence treatment effects. No previous meta analyses of psychotherapy effectiveness have been able to employ multivariable moderator analyses. The final aim was to assess the quality of each meta-analytic comparison using the Grading of Recommendations, Assessment, Development and Evaluations (GRADE, Guyatt et al., 2008) process.

Method

Search Strategy and Eligibility

A systematic review and meta-analysis was conducted using the Preferred Reporting Items for Systematic Review and Meta-Analysis guidelines (PRISMA, Moher et al., 2009) following pre-registration on PROSPERO (CRD42020175235).

Inclusion and exclusion criteria are summarized in Table 1 using the PICOS framework (population, intervention, comparator, outcome, setting). Three electronic databases (MEDLINE, CINAHL and PsycInfo) were searched for studies using a pre-developed list of key terms. Terms were selected based on their use in prior reviews of psychotherapy effectiveness (Cahill et al., 2010; Stewart & Chambless, 2009, Appendix A). For inclusion in the current review studies were required to have a methodologically and psychologically relevant term in the title or abstract. Psychological relevance was set using 'psycho-' (for MEDLINE and CINAHL) or 'psycho-'/'therap-' (PsycInfo). Use of the 'therap-' in MEDLINE and CINAHL produced an unmanageable number of irrelevant hits and was subsequently removed. Limiters included 'adult population' and 'English language' for all available studies (- April 2020). No exclusions were made based on the type of publication.

Studies were required to have included psychological therapy as conducted in a routine/naturalistic setting (i.e. locations where patients are ordinarily seen for therapy, typical referral procedures). Studies were anticipated to be of an observational nature (i.e., open/pilot trials, case series, audit/service evaluation, benchmarking).

Participants

Samples were required to be exclusively adult (Aged 16≥). If the age range for the study sample fell below 16 then the sample was excluded. Treatments could be for psychological disorders or physical health conditions which are associated with psychological distress. No exclusions were imposed regarding diagnosis/presenting problem.

Table 1.1
Inclusion and exclusion criteria used in the current review shown using the PICOS framework (population, intervention, comparator, outcome, setting).

Criteria	Inclusion	Exclusion
Population	Sample exclusively aged 16 and above (lower end of sample age range is at least 16).	Adolescent/child samples with a lower age limit below 16.
Intervention	Psychological intervention which includes individual face-to-face psycholgoical therapy (i.e. at least one session).	Samples which indicate that any proportion of patients did not recieve at least one session of individual psychological therapy.
Comparator	Studies with pre and post intervention time points. Post intervention defined here as up to six months following treatment.	(i) Studies which do not report both pre and post intervention time points. (ii) Studies for which the post intervention time point is beyond six months following treatment termination. (iii) Treatment randomisation proceadures.
Outcome	Psychological treatment effectiveness using a validated self-report measurement tool.	Service/settings which do not use a self-report measure of psychological effectiveness. Clinician reported measures were not included in this review.
Setting	Services for which a patient could expect to access psychological therapy (i.e. routine services).	Service/settings that strongly do not appear naturalistic or reflect routine practice.
	(i) Pre-post treatment designs. (ii) Studies which do not use a control condition.	(i) Studies which include a control group. (iii) Studies with N = <6. (iv) Results not available/published in English.

Interventions

Psychological interventions were required to have included a component of individual (i.e. one-to-one) psychological therapy. Study samples which included any proportion of patients who had not received individual psychological therapy (e.g. only group treatment, couples counseling, family therapy) were excluded. Multi-modal treatments (e.g. inpatient treatment program, DBT) which included a component of indi-

vidual psychological therapy were included.

Comparisons and Multiple Samples

Our main outcome of interest was pre-post change for the acute-phase of treatment (outcome measured at treatment termination). Studies which included both pre and post intervention measurement points were included. Post-intervention is defined here as the last session of treatment. For studies which only recorded the post-treatment score at a latter time point (i.e. follow-up), this was coded as the post-intervention score if it was within the first 6-months following treatment ending. Studies with post-treatment measurement beyond the 6-month post-treatment time point were excluded, as this constituted longer-term effects rather than acute-phase effects.

Practice-based studies which employed randomisation/control groups, although offering greater internal validity, are not typical of routine services, and when used in combination with observational practice-based studies pose various methodological and ethical dilemmas (Nordmo et al., 2020). Because of this, various patient sub-groups (e.g. highly distressed patients) are likely to be under represented in studies which use a control group (Philips & Falkenström, in press). For this reason studies which used random allocation or active control groups were excluded.

As a number of included studies reported multiple samples, a standard procedure was developed to support sample extraction. If a pooled study sample was reported then this sample alone was extracted. If only study sub-samples were reported (e.g. CBT vs. behavioural therapy) then each sample was extracted separately if and only if they were independent from each other (i.e., the same patients did not appear in both samples). The exception to this was when both completer and intention-to-treat (ITT) samples were reported, in which case the ITT sample only was extracted. When multiple studies used the same or overlapping data-sets then only one was study was included. The decision of which study to exclude was made by the first author on a case-by-case basis.

Outcomes

The outcomes of interest were patient-reported pre-post treatment measures. Outcome studies which employed clinician-rated measures were excluded to reduce heterogeneity. Only validated outcome measures that assessed psychotherapy effectiveness (i.e. not process, predictors, well-being or satisfaction) were included. Three broad outcome domains were employed, for which each study sample could contribute up to one measure. These domains included depression measures (e.g. Beck's Depression Inventory [BDI], Beck et al., 1996), anxiety measures (e.g. Generalised Anxiety Disorder Scale [GAD-7], Spitzer et al., 2006) and general measures of psychological distress and functioning that did not specifically measure depression or anxiety (this broad category could include measures of other forms of distress, such as symptoms of obsessive-compulsive, post-traumatic stress disorder, etc.). For samples which reported multiple measures appropriate for a single outcome domain then a preference system was followed (Appendix B). This system gave priority to global functioning measures (e.g. Outcome Questionnaire-45 [OQ-45], Lambert et al., 2004) and measures which are more frequently used across routine services.

Study Selection

Search results were exported from electronic databases (.ris files) and imported to reference management software (Mendeley, Zotero) for removal of duplicates. Unique results were imported to a web-based program for title/abstract screening using a data-mining approach ('Rayyan,' Ouzzani et al., 2016). All search results were individually screened by the first author using a pre-developed and piloted screening tool (Appendix C). This was performed for title/abstracts, and then full-texts. A sub-sample of articles were screened by a second rater at each screening stage. This included 20% of titles/abstracts (by a trainee clinical psychologist) and 10% of full-texts (by a qualified clinical psychologist). Agreement and inter-rater reliability statistics (Kappa [κ], Cohen, 1960) were used to quantify screening precision. Descriptive classifiers available for interpreting κ were employed (Landis & Koch, 1977), 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). There was substantial reliability ($\kappa = 0.78$) shown at the abstract/title screening stage (1713/1740, 98.45%) and strong reliability ($\kappa = 0.65$) at the full-text screening stage (24/30, 80%).

Additional Papers

Full-text manuscripts from the electronic database search which progressed to data extraction received two additional checks. First, reference lists were scanned for relevant article titles (i.e. backwards reference searching). Second, studies which cited the included articles were scanned (using GoogleScholar) for relevant titles (i.e. forward citation searching). For grey literature a pragmatic search was conducted using GoogleScholar (terms = "psychotherapy," AND "routine practice" AND "effectiveness") and reviewing the first 50 pages of results.

For the vast majority of studies included in the narrative review (212/252, 84.13%), corresponding authors were contacted via e-mail for additional effect-size information (see 'Effect-size calculation') with a two-week response time. Within the same e-mail, authors were invited to provide/recommend additional papers which they perceived as relevant to the review. This invitation was only performed for studies identified through the electronic database search. Of the information requests made 177 were for authors to provide correlations while 35 were for additional data (e.g. M, SD etc.). E-mail responses were received for 76 authors (35.85%). A number of these authors informed us that they did not have access to the data. Data was provided from authors for 41 samples (19.34%).

Extraction and Coding

Data extraction was performed in two phases. First, studies from the systematic database search, and second for all additional studies. To support the data extraction process a standardised extraction sheet was developed using Microsoft Excel. This spreadsheet was tested with a sample of studies (k = 10), and peer-reviewed by the research team. A coding sheet, summarised in Table 2, was developed to provide a uni-

form system for defining the levels of each moderator variable. Data from a sub-sample of manuscripts (n = 29) were extracted by a second extractor which demonstrated almost perfect reliability ($\kappa = 0.97$, agreement = 97.56%).

Sample Characteristics. There was high variability of demographic reporting for each study (e.g. gender, age, ethnicity etc.). For demographic information, the (i) mean age of each sample was extracted, and then (when reported) the number and percentage of: (ii) female, (iii), minority ethnic group, (iv) full-time employed, (v) and married patients. Each of these variables were summarised by averaging across mean averages for studies which reported this information.

Methodological Information. For methodological information the type of completion analysis used was extracted. Samples were coded as either true ITT (everyone had an equal chance of inclusion), modified ITT (when ITT was applied following prior conditions, e.g. "all who patient who attended three sessions was included"), or completers. The stage of the hour-glass model was also recorded for each effectiveness study. Samples were rated as either stage-1 (pilot and preliminary effectiveness studies) or stage-3 (evaluation/benchmarking studies studies). The region (country and continent) was recorded; studies from the UK were separated from mainland Europe, due to the high volume of effectiveness research originating in the UK.

Service Information. The type of service and associated sector were extracted for each study. As there were a large number of different sectors represented, a grouping system clustered similar sectors together. Services from primary care, health settings, counseling, and voluntary services were collated into a 'primary' sector category. Services delivering interventions for more specialist, complex or enduring presentations were grouped into a 'secondary' category. This included specialist/tertiary therapy services/clinics, community mental health teams/centers, and intensive out-patient services. University based services (either training clinics or counseling centers) were assigned to a 'University clinics' category. Finally, inpatient, day hospital and partial

Table 1.2
Sumary coding sheet for extracting study information and categorising by level of moderator sub-group. These moderators form the categorical, sub-group variables for the current study.

Moderator	Level	Description
Setting	Outpatient	Sample of patients treated at an out-patient settings.
	Inpatient	Sample of patients treated at either an (i) inpatient; (ii) day hospital; (iii) residential; or (iv) partial hospital setting.
Completion	Completer	Sample of patients who all completed treatment.
	ITT	Sample of patients who used intention-to-treat principles. This is either (i) true ITT; or (ii) modified ITT (i.e. a minimum number of attended sessions).
Sector	University Clinics	Sample of patients seen at (i) University training clinics; or (ii) University based out-patient clinics.
	Primary	Sample of patients seen at a: (i) primary care; (ii) health; (iii) counselling/University counselling; (iv) voluntary ; (v) private [independent or group]; or (vi) employee assistential/occupational health service.
	Secondary	Sample of patients seen at a: (i) secondary care; (ii) CMHTs /CMHC; (iii) tertiary/specialised psychotherapy; (iv) behavioural health/managed care ; or (v) Intensive out-patient setting.
	Inpatient	Sample of patients treated at either an (i) inpatient; (ii) day hospital; (iii) residential; or (iv) partial hospital setting.
Continent	Continents	Continent of study setting, consisting of either: (i) UK; (ii) mainland Europe; (iii) North America; (iv) Asia; (v) Australasia.
Therapy	Dynamic	Therapy or counselling which follows apsychodynamic orientation.
	CBT	Therapy or counselling which follows a cognitive and/or behavioural orientation.
	Counselling	Counselling which is either (i) person-centered; or (ii) orientation not specified.
	Other	Therapy or counselling which (i) has not been mentioned above, or (ii) is not specified/reported in the study manuscript.
Trainee	Unqualified	Interventions exclusively made up of psychology trainees.
	Other	All other samples/studies.
Hour-glass	Stage-1	Methodologies including: (i) pilot or (ii) preliminary effectiveness studies.
	Stage-3	Methodologies including: (i) service evaluatons; (ii) benchmarking; (iii) routine outcome reporting; (iv) predictors of outcome/drop-out.

hospital services were grouped into a 'inpatient' category. Whether or not study interventionists consisted of clinicians in training was also recorded as a separate variable. We defined clinicians in training as staff training towards a professional psychology training course (i.e. clinical psychology interns/students, training psychiatrists or assistant psychologists). Staff who were not psychologists or qualified therapists, but who had a core profession (e.g. nurses, social workers) were not recognised as unqualified interventionists.

Treatment Information. The treatment delivered was recorded for each study. Treatments were then assigned to a broad meta-therapy category, including: (i) cognitive and/or behavioural, (ii) dynamic/interpersonal, (iii) person-centered counseling (or counseling without a specified orientation), or (iv) other/non-specified. The average number of sessions was also recorded. For studies that reported the mean number of sessions then this was the metric extracted. For studies that alternatively used a time metric (days/weeks/months/years) then a uniform metric was applied (i.e. conversion to days). There was subsequently two possible dosage metrics, sessions of treatment and treatment days. If studies reported sample dosage, but with an alternative measure of central tendency (i.e. median) then this was converted to mean average.

Effect-Size Calculation. All analyses were conducted using the R statistical analysis environment (R Core Team, 2020, v 4.0.2). Effect-size calculation and meta-analyses were performed using the metafor (Viechtbauer, 2020), dmetar (Harrer et al., 2019a), and meta (Schwarzer, 2020) packages.

Paired-samples Cohen's d (Standard mean change, Cohen, 1988) was computed for each study sample by dividing the pre-post mean change by the pre-treatment standard deviation (see Figure 1). Sample variance was adjusted using Pearson's r in order to account for the inherent violation of independence (i.e. regression to the mean) in pre-post comparisons (Cuijpers et al., 2017). This approach has been advocated for benchmarking of pre-post outcomes studies (Minami et al., 2008).

$$d = \frac{Mean^2 - Mean^1}{SD^1}$$

Figure 1.1: Formula for standardised mean change. Where 1 is the sample pre-intervention and 2 is the sample following intervention.

Due to the fact that the majority of manuscripts did not report all of the required information for calculating this variant of Cohen's d, a hierarchical stepped approach was used to handle the missing information (see Table 3). For studies which reported all of the required information (N, M^1, M^2, SD^1, r) then d was calculated without additional consideration. For studies which (commonly) did not report r then we e-mailed corresponding authors (two-week response time) to request this information. When unsuccessful r was imputed using an empirically supported estimate (r = .60, Balk et al., 2012). For studies which did not provide the more fundamental figures (M^1, M^2 or SD^1) but reported a paired samples Cohen's d (any variant) then this effect-size was extracted. For studies which did not report fundamental figures and did not report a Cohen's d, then e-mail requests were sent to corresponding authors. If this was unsuccessful then we applied conversion formulas in situations when studies reported alternative quantitative metrics (e.g. median, range, standard error, ANOVA, regression) to generate means and standard deviations. In situations when all of these steps were unsuccessful/not applicable then the studies in question were removed from the meta-analyses (included in the narrative synthesis only).

Risk of Bias and Methodological Quality Assessment

The Joanna Briggs Institute Quality Appraisal Tool for Case Series (Appendix D, Munn et al., 2020) was used to assess risk of bias for all reviewed studies. The items within this tool were judged by the review team to be of relevance to studies in naturalistic settings. Two items concerning outcome measurement were removed to make an adapted 8-item tool. This is because the review inclusion criteria (validated nomothetic measure of effectiveness) would implicitly mean that every study would

Table 1.3
Hierarchical proceadure for effect-size calculation.

Steps	Scenario	Response
Step 1	Manuscript reports all required information (N, M1, M2, SD1) for preferred d.	Calculate preferred d.
Step 2	Manuscript reports all information apart from Pearson's r.	E-mail corresponding authors to request r.
Step 3	Manuscript does not report the mean or standard deviation but reports paired samples d.	Use the reported d within the manuscript.
Step 4	Manuscript does not report mean, standard deviation, or paired samples d however reports alternative metrics (e.g. median, range, standard error, ANOVA, regression	Estimate the meand and standard deviation by converting available metrics.
Step 5	All above steps attempted without success	Study is not included in meta-analysis but is retained for narrative synthesis.

meet these criteria. Included study samples were rated as having either met or not met each criteria (yes/no/not sure). The authors of the tools do not provide a scoring classification system or cut-off (Munn et al., 2020). They advise that such decisions should be made by the reviewers who employ them. For this review each 'yes' was given a score of one. Each study subsequently received a cumulative bias score (range = 0-8) with higher scores indicating less risk of bias. All studies were rated by the first author while a sub-sample (23.8%) was second rated by a pair of MSc psychological research methods students (11.9% each). Inter-rater reliability at this stage was substantial (κ = 0.67, agreement = 86.25%)

The methodological quality of the evidence within each meta-analytic comparison was assessed by three reviewers using guidelines for the Grading of Recommendations, Assessment, Development and Evaluations (GRADE, Guyatt et al., 2008). 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: (i) risk of bias within included studies, (ii) inconsistencies in aggregated treatment

effect, (iii) indirectness of evidence, (iv) imprecision, and (v) publication bias.

Data Synthesis

Random-effects meta-analyses were used to estimate pooled effect sizes. Pooled and weighted effect-sizes with 95% confidence intervals were calculated for all included study samples. Due to the anticipated high number of studies, forest plots without study details were employed to illustrate the pattern of study effects and the overall pooled estimate. This decision was made to aid visual interpretation as the large volume of included studies meant identifying individual study effect sizes from the plot was challenging (individual study effect sizes are reported within the Supplementary Material). The number of patients needed to treat (i.e. number needed to treat, NNT) in order for one patient to receive a positive outcome was calculated using the method proposed by Kraemer & Kupfer (2006).

The extent of between-study heterogeneity was assessed using I^2 (Higgins & Thompson, 2002) and the Q statistic (Cochran, 1954). I^2 was interpreted as low (25-50%), moderate (50-75%) or high (75-100%, Higgins et al., 2003). The impact of publication bias on treatment estimates was visualised using funnel plots and assessed statistically using rank correlation tests (Begg & Mazumdar, 1994), Egger's regression test for funnel plot asymmetry (Egger et al., 1997), and fail-safe N (Rosenthal method, Rosenthal, 1979).

Moderator Analyses

Pre-defined moderator analyses were conducted for sub-groups, distinguished by categorical variables while meta-regression was used for continuous variables. There were seven sub-group moderator variables: (i) setting, (ii) type of completion sample, (iii) sector, (iv) region, (v) therapy modality, (vi) experience, and (vii) stage of the treatment evaluation (i.e. hour-glass model). There was eight continuous moderator variables. These included: (i) year of study publication, (ii) average age of sample, (iii) treatment dosage (i.e. number of out-patient sessions), (iv) rate of sample from a minority ethnic background, (v) rate of sample married, (vi) rate of sample in full-time

employment, (vii) rate of sample who were female, and (viii) study risk of bias scores (arising from risk of bias assessment). Bonferroni adjustments were applied to each group of moderators, resulting in p-values of .00714 for categorical variables (.05/7) and .00625 for continuous variables (.05/8).

For moderators that produced significant meta-analytic models then between sub-group pairwise comparisons were made. This consisted of inspecting whether sub-group effect-size classifications differed and also whether confidence intervals showed overlap.

The approach to multivariable moderator analysis followed available guidance (Harrer et al., 2019b). Variables were selected for multi-variate analysis based on suspected interactions by the research team. This included completion methodology and mean number of sessions (in influencing effect-size was assessed). Completion analysis, coded as a dummy variable (ITT vs. completion) was first entered into an initial multi-regressive model. A second 'full' model was then built using both completion analysis and mean number of sessions. Differences in these two models were assessed using a log-likelihood ratio rest, with a p-value of p = <.05 required to indicate significant model differences. If models were significantly different then comparisons were made between log-likelihood scores and information criteria statistics. A lower log-likelihood score, and smaller Akaike's-information criteria (AIC) score would indicate improved model fit for the full model. Following this, completion methodology and treatment dosage were modeled as interaction terms in the final model. This stage was conducted regardless of whether predictors were significant in isolation within previous models. If interaction models were significant then coefficient interpretations were made.

Results

Search Results

The systematic search of electronic databases produced 10,503 results. After removal of duplicates, this was reduced to 8,709. Following title/abstract screening there were 325 potentially eligible records remaining. Of these, 30 manuscripts were not available through the first author's institution. E-mails were sent to corresponding authors to request access. This led to the retrieval of a further 8 manuscripts. Articles which remained without full-text manuscripts (22, 6.77%) were excluded from the review on the basis that the full-text was not available for eligibility screening. On completion of the full-text screening process there were 130 studies remaining. All of these articles were included in the narrative (qualitative) synthesis.

Through forward citation searching and backwards reference searching a further 197 articles were identified. Finally, 97 articles were found through grey literature/provided by authors. A PRISMA flow diagram (Schulz et al., 2010) is shown in Figure 2. A break-down of screening decision results are provided in Table 4 with a full-list in appendix E.

Table 1.4
Full-text screening decision results for all studies

Decision	Phase 1	Phase 2	Phase 3	Total
Exclude	174	111	54	339
Include	130	79	42	251
No Access	21	7	1	29
Total	325	197	97	619

Note. Difinitions for different phases are as follows:
Phase 1 = studies identified through the electonic database search.
Phase 2 = studies identified through phase 1 reference lists and citation searching.
Phase 3 = grey literature and studies provided by authors contacted during the study.

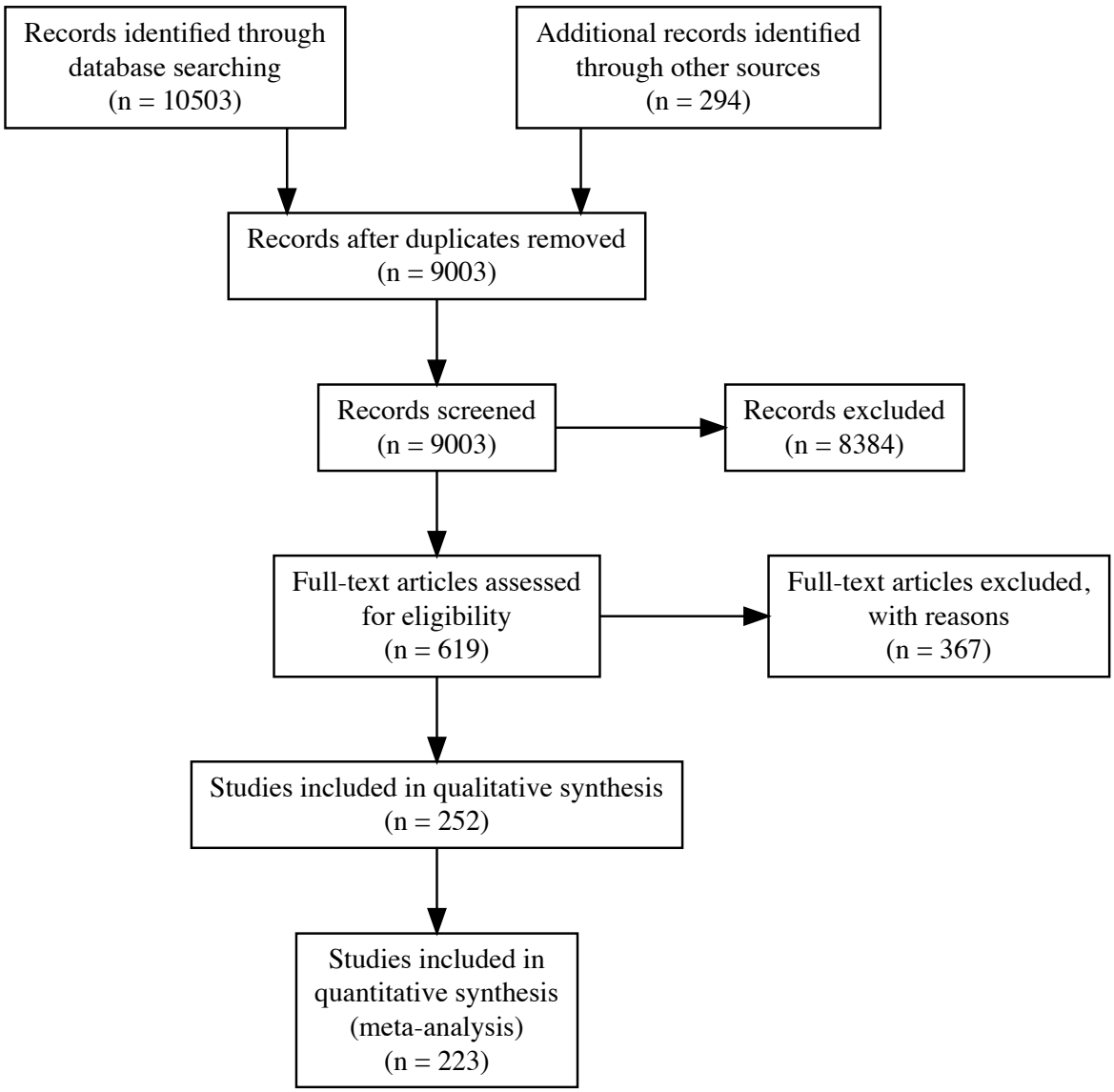

Figure 1.2: Prisma flow diagram of studies throughout the review.

Narrative Synthesis

All 252 studies included in the data extraction phase were included in the narrative synthesis. Of these studies, 223 were eligible for inclusion in the meta-analysis. As a number of studies provided multiple samples, the total number of samples exceeded the number of studies. There was subsequently 298 samples in the narrative synthesis, while there was 263 samples included in the meta-analysis. Full details of the study characteristics of all included studies are reported in the Supplementary Material 1, while summary statistics are provided in Table 5.

Methodological Information

Of the samples included in the narrative synthesis, the year of publication ranged from 1984 to 2020. The median publication date was 2013. There were 294 samples reported since 2000, 213 since 2010 and 126 since 2015.

The number of ITT samples, including those that used modified ITT (i.e. studies specifying a minimum number of attended sessions for inclusion) or when ITT could only be assumed was 169. The number of studies that used our more rigorous definition of ITT was 64, while the number of studies that used a completer sample was 118. When distinguishing studies based on the stage of the hour-glass model there was 34 (11.41%) samples at stage-1 and 264 (88.59%) at stage three.

Sample Characteristics

Demographic information was reported for 291 of the 298 samples included in the narrative synthesis. The demographic sample size for samples included in the narrative synthesis ranged from 4 (Sauer-Zavala et al., 2019) to 33,243 (Pybis et al., 2017, CBT sample). The pooled demographic sample size for the narrative synthesis was 233,140. Self-reported gender information was available for the majority of samples (k = 279). Of these samples 144,273 (61.88%) patients were females. When averaging across available percentages, there was a pooled average of 66.0% females. There were 13 exclusively female samples and 2 exclusively male samples. Within studies that

reported a mean average age the pooled average age was 35.33 years (range = 19.00 - 60.50).

There were 127 samples that reported the number or percentage of patients from minority ethnic backgrounds. The mean percentage of patients from minority ethnic backgrounds was 23.00%. For marital status, 106 samples reported relevant data with a mean average (patients who were married) of 23.00%. There were 96 samples that reported employment status. The mean percentage of patients in employment across samples was 56.00%.

Table 1.5
Summary statistics across the pooled sample and also by sector for varying variables.

	Level	Uni Clin	Primary	Secondary	Inpatient	Other	Total
N	Female	5350	95373	14952	5797	22801	144273
	Total	9195	158150	22586	9515	33694	233140
Age	samples	65	77	82	29	7	260
	mean	33	36	35	34	36	176
	min	20	19	21	24	24	109
	max	52	60	52	47	46	258
Sessions	samples	54	64	54	4	6	182
	mean	21	11	14	13	8	69
	min	2	4	1	9	8	24
	max	85	64	64	24	9	247
Setting	Mixed	0	0	0	0	5	5
	Outpatient	68	96	91	0	4	259
	Inpatient	0	0	1	33	0	34
Completion	ITT	48	48	53	16	4	169
	Check	1	2	4	1	1	9
	Completers	19	45	35	16	3	118
Therapy	CBT	43	41	49	14	5	152
	Counselling	0	22	3	0	0	25
	Dynamic	12	9	16	13	0	50
	Other	13	24	24	6	4	71
Hour Glass	Stage-1	4	6	16	7	1	34
	Stage-3	64	90	76	26	8	264
Continent	Asia	4	1	0	0	1	6
	Australasia	5	0	5	0	0	10
	Europe	20	13	14	15	1	63
	N.America	38	32	39	10	4	123
	UK	1	50	34	8	3	96
Continent	Total	68	96	92	33	9	298

Service Information

The country that contributed most samples was the USA (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, when differentiating the UK from mainland Europe, the order of continental representation was North America (k = 123), the UK (k = 96), mainland Europe (k = 63), Australasia (k = 10), and Asia (k = 6).

For treatment setting, there was 96 (32.21%) samples were in the primary care category, 92 (30.87%) in the secondary care category, 33 (30.87%) in the inpatient care category, and 68 (22.82%) from University clinics. There was 9 (3.02%) samples from a combination of sectors (i.e. other)

Treatment Information

In terms of treatment modality, samples received 152 treatments classified as cognitive and/or behavioural therapies, 50 received dynamic/interpersonal therapy, 25 were classified as non-specific or person-centered counseling, and 71 classified as other. Treatment duration metrics were reported for the majority of study samples (k = 256). The most common treatment duration metric was sessions (or hours, k = 225), followed by months (k = 12), and then days (k = 8). There was no treatment duration metric available for 42 samples. The pooled mean across those studies which reported the mean number of sessions, was 16.30 sessions (range = 1.00-139.30).

There were 62 62 samples reported as exclusively consisting of unqualified (i.e., trainee) clinicians; while 100 samples reported having at least one unqualified clinician.

Risk of Bias and Methodological Quality

In terms of study risk of bias assessment, mean average bias score across samples included in the narrative review was 5.53 (SD = 1.47, range = 1-8). Total bias score for each individual study is reported in Supplementary Material 1. The most frequently met criteria was for demographic reporting detail (264/298), followed by

service reporting detail (260/298). The least frequently met criteria was for complete inclusion (i.e. consecutive recruitment and intention-to-treat analysis, 41/298) followed by consecutive inclusion (93/298).

Meta-Analyses

Each outcome domain had a primary meta-analysis. A summary of the primary meta-analyses is shown in Table 6. There was a wide variety of specific measures employed within each meta-analysis (as shown in Appendix B). 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 this overall rating was reduced in level to 'very low' based on study limitations and also inconsistency within the available evidence.

Table 1.6
Findings from the primary meta-analyses.

Variable	k	ES	Lower	Upper	p	I2	Q
Depression	140	0.98	0.90	1.06	< 0.001	98.40	3037.46
Anxiety	84	0.83	0.73	0.92	< 0.001	97.52	1488.88
General	184	1.01	0.93	1.08	< 0.001	98.92	15685.18

Depression

For depression outcomes, (k = 140 samples), 10 different outcome measures were used. The most frequently used depression measure was the Beck Depression Inventory (BDI I or II, k = 78), followed by the Patient Health Questionnaire (PHQ-9, k = 30) and then the Brief Symptom Inventory ([BSI] Depression Index, k = 8). The depression meta-analysis had a combined N of 68,077. Individual study effect-sizes are illustrated in the depression forestplot in Figure 3. The pooled effect-size was significant, indicative of a large (d = 0.98, [CI 0.9-1.06], p = < 0.001, GRADE = very low) reduction in depression symptoms. The number of patients needed-to-treat in order to provide one patient with a positive outcome was 1.95. There was evidence of significant study heterogeneity (I^2 = 98.4%, Q[df = 139] = 3,037.46, p = < 0.001). The funnel plot in Figure 4 shows limited visual evidence of asymmetry. The funnel

rank correlation test was not significant ($\tau = 0.028$, p = 0.629). In contrast, the funnel regression test was significant (Z = 2.665, p = 0.0077). The fail-safe N indicating the number of studies reporting no intervention effect that would be required to make the aggregated effect not significant was N=736,945.

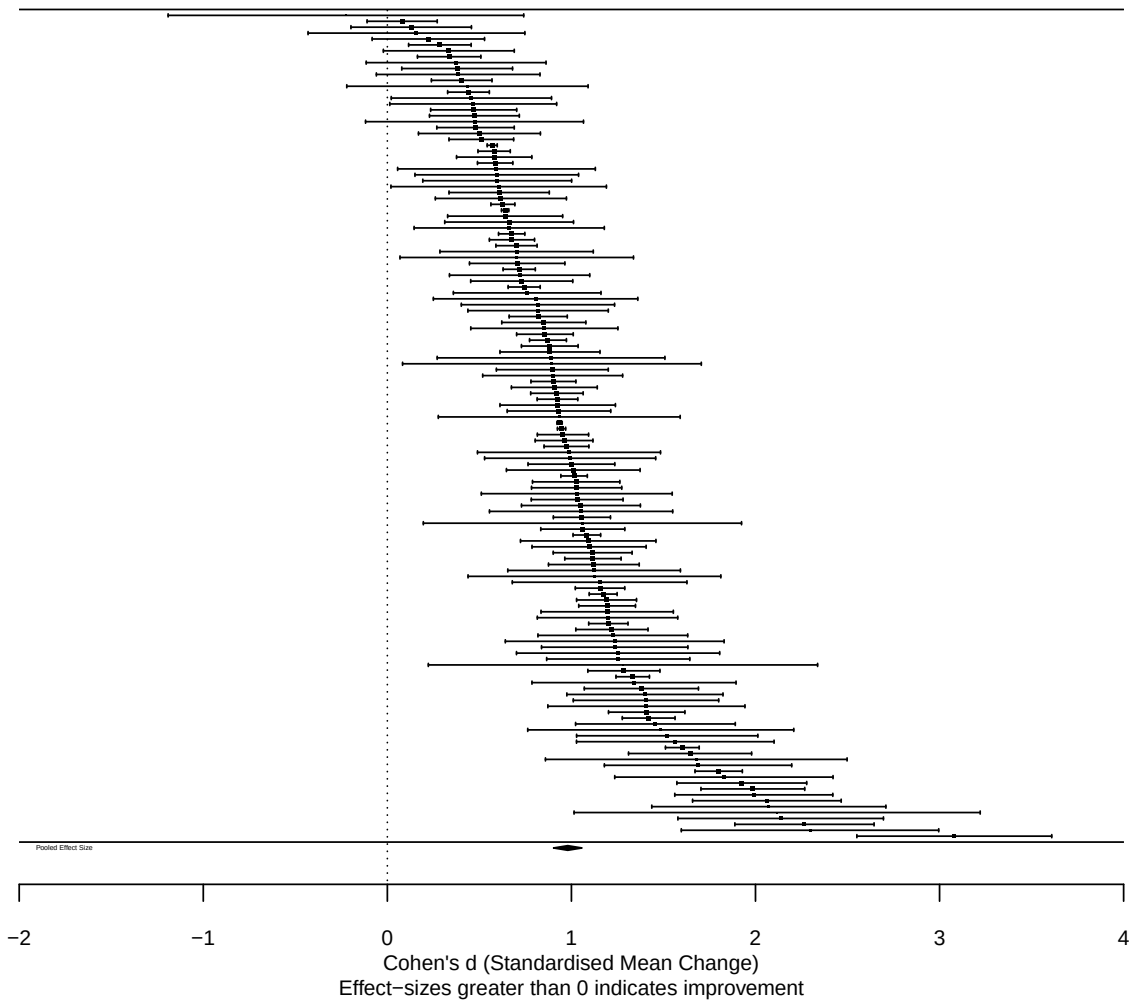

Figure 1.3: forestplot of pre-post psychological therapy effect sizes for depression outcomes.
Square boxes depict individual study Cohen's d effect sizes, error bars display 95 percent confidence intervals and the diamond represents the pooled estimate effect.

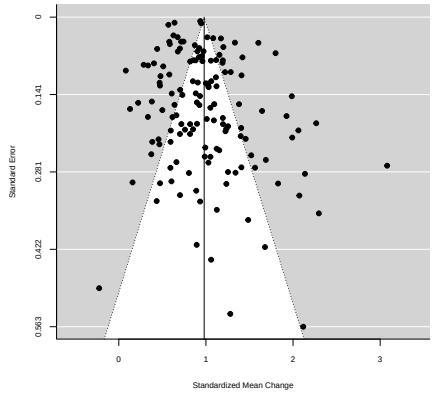

Figure 1.4: Funnel plot of the distribution of studies reporting pre-post depression outcomes.

Anxiety

For anxiety outcomes, (k = 84 samples), 20 different outcome measures were used. The most frequently used measure was the the Beck Anxiety Inventory (BAI, k = 19), followed by the Generalised Anxiety Disorder (GAD-7, k = 19), and then the Brief Symptom Inventory ([BSI] Anxiety Index, k = 8). The anxiety meta-analysis had a combined N of 26,689. Individual study effect-sizes are illustrated in the anxiety forestplot in Figure 5. The pooled effect-size was significant, indicative of a large (d = 0.83, [CI 0.73-0.92], p = < 0.001, GRADE = very low) reduction in anxiety symptoms. The number of patients needed-to-treat in order to provide one patient with a positive outcome was 2.26. There was evidence of significant study heterogeneity (I^2 = 97.52%, Q[df = 83] = 1,488.88, p = < 0.001). The funnel plot in Figure 6 shows limited evidence of asymmetry. The funnel rank correlation test was not significant (τ = 0.061, p = 0.416). In contrast, the funnel regression test was significant (Z = 3.186, p = 0.0014). The fail-safe N was 155,478

General

For general outcomes, (k = 184 samples), 40 different measures were used. The most frequently used measure was the CORE (10/OM, k = 40), followed by the Brief Symptom Inventory ([BSI] Global Severity Index, k = 26), the Symptom Checklist 90 ([SCL] Global Severity Index, k = 21) and then the Outcome Questionnaire (OQ 30/45/45.2, k = 14). The meta-analysis for general outcomes had a combined N of 126,734. Individual study effect-sizes are illustrated in the general forestplot in Figure 7. The pooled effect-size was significant, indicative of a large (d = 1.01, [CI 0.93-1.08], p = < 0.001, GRADE = very low) reduction in general symptoms. The number of patients needed-to-treat in order to provide one patient with a positive outcome was 1.91. There was evidence of significant study heterogeneity across the included studies (I^2 = 98.92%, Q[df = 183] = 15,685.18, p = < 0.001). The funnel plot (see Figure 8) shows a degree of asymmetry with clustering to the right of the mid-line. The funnel rank correlation test was significant (τ = 0.228, p = <0.001). In contrast, the funnel

regression test was not significant (Z = -0.733, p = 0.46). The fail-safe N was 2,018,805.

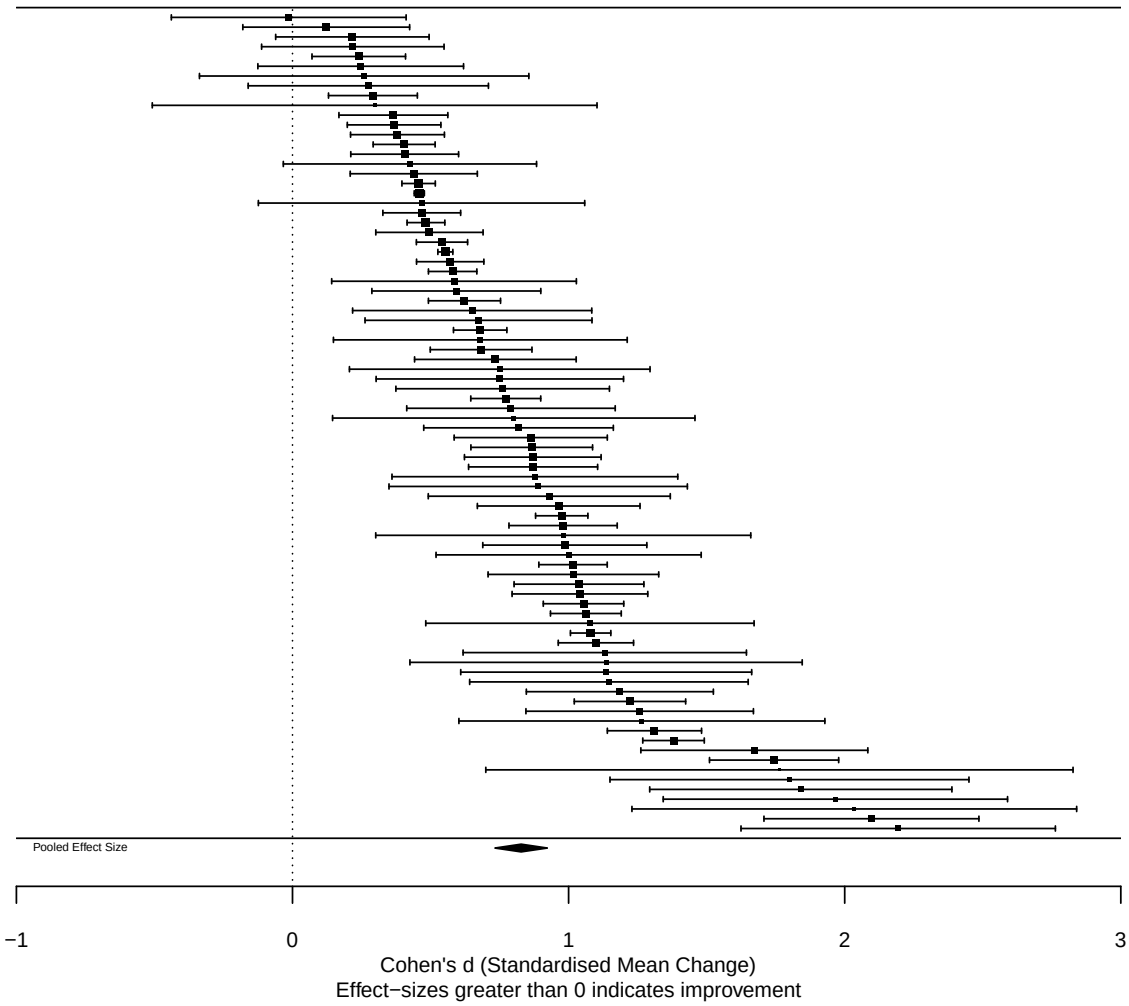

Figure 1.5: forestplot of pre-post psychological therapy effect sizes for anxiety outcomes.
Square boxes depict individual study Cohen's d effect sizes, error bars display 95 percent confidence intervals and the diamond represents the pooled estimate effect.

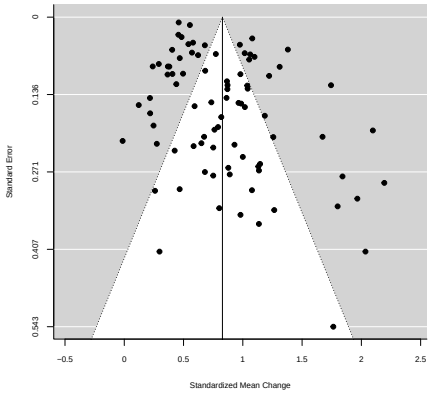

Figure 1.6: Funnel plot of the distribution of studies reporting pre-post anxiety outcomes.

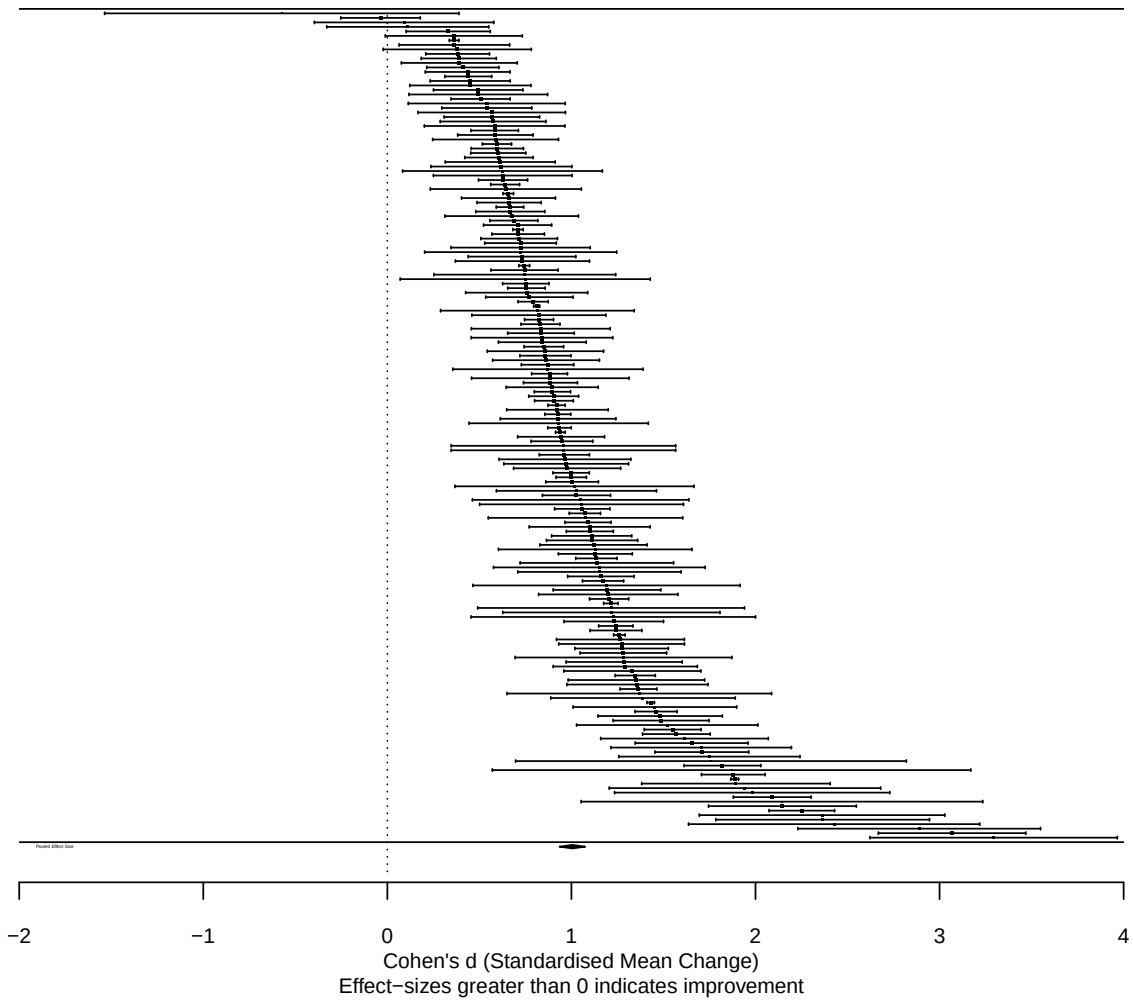

Figure 1.7: forestplot of pre-post psychological therapy effect sizes for general outcomes.
Square boxes depict individual study Cohen's d effect sizes, error bars display 95 percent confidence intervals and the diamond represents the pooled estimate effect.

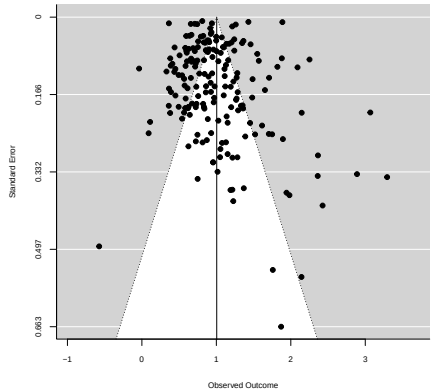

Figure 1.8: Funnel plot of the distribution of studies reporting pre-post general outcomes.

Moderator Analyses

Univariate Moderators. Categorical moderator analyses (i.e. sub-groups) are reported in Tables 7-9. Of the eight moderators, there were two variables that were significant for all three outcome domains (completion sample and continent). Completer analyses consistently had larger effect sizes compared to ITT analyses across all samples, with no overlap between confidence intervals (CI). For continent, UK and North American studies had larger effect sizes than other continents (mainland Europe, Australasia and Asia), with varying levels of overlap in CI among the other subgroups. To a lesser extent Europe also had larger effect sizes than Australasia and Asia. UK and North American pooled effect-sizes were comparable to each other across domains.

The remaining six moderators were not consistent across outcome domains. Study setting was significant for the anxiety and general domains, although both analyses showed CI overlap. Outpatient samples out-performed inpatient care for anxiety, while the reverse was shown for the general domain. Sector was significant for the anxiety and general domains. Anxiety samples showed greater average outcome within primary and University clinic sectors. For general outcomes, secondary services and University clinics had lower effect sizes than other sectors. Type of therapy was significant for the anxiety and general domains. Anxiety samples which accessed dynamic based therapies had larger effect sizes compared to other interventions (with overlapping CI). For general samples, CBT-based interventions had higher effect sizes compared to the other therapy meta-categories with no CI overlap. Stage of the hour-glass model was significant only for the general domain. Stage-one samples (pilot studies, preliminary evaluations of treatments) had larger effect sizes than stage-three samples, with large overlap in CI. Finally, for experience (i.e. exclusively unqualified samples vs. not), significant results were found for the anxiety and general domains. Anxiety samples exclusively consisting of unqualified clinicians had higher effect sizes than other samples (no CI overlap). For the general domain the reverse was shown, unqualified clinician samples had lower effect sizes than other samples (no CI overlap).

Between-study heterogeneity was also explored using eight continuous variables (see table 10). Neither mean age, proportion of ethnic minority patients, or proportion of married patients were significant for any of the outcome domains. Risk of bias score was significant for all three domains, with higher quality scores linked to larger effects. Year of publication was significant only for anxiety, suggesting that more recent studies produce greater effect-sizes for anxiety. Mean number of sessions was significant only for depression, suggesting that treatment effectiveness increases in line with a greater number of sessions received. Employment was significant only for anxiety, suggesting that studies with greater employment rates show larger effect-sizes. Proportion of female patients was significant for the anxiety domain. Greater female representation was linked with lower anxiety effect-sizes.

Multivariable Moderators. Multivariable moderator analysis was conducted for completion methodology and mean number of sessions. For depression, the full model found completion methodology, but not mean number of sessions to be a significant individual predictor. The overall test of moderators was significant. There was no significant difference in model fit, based on the log-likelihood ratio test. For the interaction model, the overall test of moderators was significant, however neither of the predictor variables or the interaction term were significant in isolation.

For anxiety, the full model found neither completion methodology or mean number of sessions to be a significant individual predictors. The overall test of moderators was not significant. There was no significant difference in model fit, based on the log-likelihood ratio test. For the interaction model, the overall test of moderators was not significant. The individual predictor variables and also the interaction term were not significant in isolation.

For general outcomes, the full model found neither completion methodology or mean number of sessions to be a significant individual predictors. The overall test of moderators was not significant. There was no significant difference in model fit, based on the log-likelihood ratio test. For the interaction model, the overall test of moderators

was not significant. The individual predictor variables and also the interaction term were not significant in isolation.

Table 1.7
Sub-group (categorical) moderator analyses for depression outcomes.

Moderator	Level	k	Effect Size	Confidence Intervals	Q	I2
Random effects model for sector (Q = 5.99, p = 0.2)						
Sector	Primary	31	1.06	0.99 - 1.13	9547771.37	1.00
	Uni. Clinics	29	0.96	0.86 - 1.07	43195.81	1.00
	Secondary	55	0.97	0.91 - 1.04	80785.29	1.00
	Inpatient	15	0.91	0.7 - 1.12	237284.01	1.00
Random effects model for ITT (Q = 13.88, p = <0.001**)						
	ITT	76	0.92	0.88 - 0.97	9728418.18	1.00
Completion	Completers	58	1.09	1.01 - 1.17	220978.90	1.00
Random effects model for setting (Q = 3.82, p = 0.148)						
	Outpatient	115	0.99	0.95 - 1.02	9657600.36	1.00
Setting	Inpatient	16	0.92	0.74 - 1.1	240002.33	1.00
Random effects model for continent (Q = 27.23, p = < 0.001**)						
Continent	N.America	56	0.99	0.9 - 1.08	640246.22	1.00
	UK	43	1.09	1.05 - 1.14	3564834.35	1.00
	Europe	26	0.94	0.82 - 1.05	59071.64	1.00
	Australasia	4	0.67	0.33 - 1	7087.67	1.00
	Asia	5	0.59	0.35 - 0.83	91.78	0.96
Random effects model for therapy modality (Q = 1.5, p = 0.682)						
Therapy	Dynamic	22	1.01	0.82 - 1.19	41858.50	1.00
	Counselling	6	0.89	0.72 - 1.07	3471906.01	1.00
	CBT	88	1.00	0.96 - 1.05	393105.73	1.00
	Other	18	0.98	0.77 - 1.19	307046.63	1.00
Random effect model for training samples (Q = 1.36, p = 0.244)						
	No/NA	116	1.01	0.97 - 1.04	9876336.39	1.00
Trainees	Yes	18	0.89	0.7 - 1.08	76274.10	1.00
Random effects model for hour-glass stage (Q = 1.37, p = 0.242)						
	Stage-3	113	1.00	0.96 - 1.03	9952776.03	1.00
HourGlass	Stage-1	21	0.93	0.82 - 1.04	464.36	0.96

Note. Model Outputs in Bold are significant at either * p = <.05.
** Bonferroni adjustment, p = <.007

Table 1.8
Sub-group (categorical) moderator analyses for anxiety outcomes.

Moderator	Level	k	Effect Size	Confidence Intervals	Q	I2
Random effects model for sector (Q = 128.47, p = < 0.001**)						
Sector	Primary	21	0.99	0.96 - 1.03	329379.72	1.00
	Secondary	24	0.62	0.55 - 0.69	30702.22	1.00
	Inpatient	8	0.59	0.31 - 0.88	108223.63	1.00
	Uni. Clinics	29	1.00	0.89 - 1.11	32067.93	1.00
Random effects model for ITT (Q = 7.55, p = 0.006*)						
	ITT	57	0.77	0.74 - 0.79	512107.17	1.00
Completion	Completers	26	0.96	0.82 - 1.09	92492.19	1.00
Random effects model for setting (Q = 5.75, p = 0.016*)						
	Outpatient	74	0.84	0.82 - 0.87	435141.60	1.00
Setting	Inpatient	9	0.58	0.37 - 0.79	157933.03	1.00
Random effects model for continent (Q = 26.72, p = < 0.001**)						
Continent	N.America	32	0.90	0.81 - 0.98	230641.72	1.00
	UK	25	0.89	0.8 - 0.98	115765.33	1.00
	Europe	19	0.79	0.67 - 0.91	27933.90	1.00
	Australasia	4	0.61	0.28 - 0.94	3781.78	1.00
	Asia	3	0.59	0.49 - 0.69	3.33	0.40
Random effects model for therapy modality (Q = 105.34, p = < 0.001**)						
Therapy	Dynamic	12	0.92	0.67 - 1.17	11879.61	1.00
	Counselling	2	0.43	0.38 - 0.49	28.37	0.96
	CBT	62	0.86	0.79 - 0.93	174811.70	1.00
	Other	7	0.74	0.47 - 1.02	153478.52	1.00
Random effects model for hour-glass stage (Q = 0.31, p = 0.579)						
	Stage-3	73	0.82	0.79 - 0.84	630326.22	1.00
HourGlass	Stage-1	10	0.87	0.69 - 1.05	639.27	0.99
Random effect model for training samples (Q = 13.8, p = < 0.001**)						
	No/NA	65	0.75	0.72 - 0.78	497245.30	1.00
Trainees	Yes	18	1.12	0.93 - 1.32	59913.29	1.00

Note. Model Outputs in Bold are significant at either * p = <.05.
** Bonferroni adjustment, p = <.007

Table 1.9
Sub-group (categorical) moderator analyses for general outcomes.

Moderator	Level	k	Effect Size	Confidence Intervals	Q	I2
Random effects model for sector (Q = 45.74, p = < 0.001**)						
Sector	Primary	54	1.10	0.99 - 1.2	39094444.62	1.00
	Secondary	58	0.88	0.86 - 0.9	101238.33	1.00
	Inpatient	24	1.07	0.98 - 1.17	65648.02	1.00
	Uni. Clinics	27	0.81	0.73 - 0.89	25376.01	1.00
Random effects model for ITT (Q = 9.79, p = 0.002**)						
	ITT	89	0.95	0.9 - 1	12412018.93	1.00
Completion	Completers	80	1.09	1.02 - 1.17	10416544.97	1.00
Random effects model for setting (Q = 6.18, p = 0.045*)						
	Outpatient	141	1.00	0.92 - 1.08	104239386.44	1.00
Setting	Inpatient	25	1.06	0.98 - 1.14	67134.33	1.00
Random effects model for continent (Q = 16.45, p = 0.002**)						
Continent	UK	60	1.03	0.89 - 1.18	92811055.56	1.00
	N.America	59	1.03	0.97 - 1.09	4599532.56	1.00
	Europe	41	0.98	0.9 - 1.06	143451.90	1.00
	Australasia	4	0.81	0.72 - 0.9	330.52	0.99
	Asia	5	0.91	0.58 - 1.23	575.15	0.99
Random effects model for therapy modality (Q = 46.9, p = < 0.001**)						
Therapy	CBT	77	1.18	1.12 - 1.23	171878.31	1.00
	Dynamic	34	0.88	0.8 - 0.96	48684.40	1.00
	Counselling	19	0.90	0.8 - 1.01	318722.49	1.00
	Other	39	0.87	0.71 - 1.02	103693471.80	1.00
Random effects model for hour-glass stage (Q = 0.15, p = 0.703)						
	Stage-1	23	1.06	0.86 - 1.26	4506.87	1.00
HourGlass	Stage-3	146	1.02	0.94 - 1.1	104370017.66	1.00
Random effect model for training samples (Q = 20.21, p = < 0.001**)						
	No/NA	144	1.07	0.99 - 1.15	104271844.65	1.00
Trainees	Yes	25	0.76	0.65 - 0.87	58036.50	1.00

Note. Model Outputs in Bold are significant at either * p = <.05.
** Bonferroni adjustment, p = <.007

Table 1.10
Meta-regression moderator variables (continuous) for depresison, anxiety and general outcome domains.

Domain	Moderator	Mean (range)	k	B	CI	SE	p	Q	R2
	Year (of publication)	(1988 - 2020)	134	0.00	-0.01 - 0	0.00	0.585	0.30	9.14
Depression	Mean age	(19-60 years; M = 36)	122	0.00	0 - 0	0.00	0.751	0.10	23.68
	Sessions (mean)	(1-46 sessions; M = 15)	83	0.01	0 - 0.01	0.00	0.008 *	7.10	18.27
	Ethnicity (% minority)	(0-66%; M = 23%)	61	-0.10	-0.38 - 0.18	0.14	0.482	0.49	0.00
	Martial status (% Married)	(0-73%; M = 35%)	53	-0.10	-0.71 - 0.5	0.31	0.736	0.11	2.36
	Employment (% full-time)	(5-100%; M = 52%)	44	0.37	-0.06 - 0.81	0.22	0.090	2.87	39.11
	Gender (% female)	(0-100%; M = 67%)	127	-0.06	-0.21 - 0.1	0.08	0.476	0.51	7.36
	Risk of Bias (1-10)	(1-8; M = 5.69)	134	0.02	0.01 - 0.04	0.01	<0.001**	13.31	70.63
	Year (of publication)	(1999 - 2020)	83	0.02	0.01 - 0.02	0.00	<0.0(**	52.51	0.00
	Mean age	(19-60 years; M = 35)	78	0.00	-0.01 - 0.01	0.00	0.664	0.19	0.00
Anxiety	Sessions (mean)	(1-46 sessions; M = 16)	52	0.00	0 - 0	0.00	0.997	0.00	0.00
	Ethnicity (% minority)	(0-59%; M = 19%)	40	0.33	-0.19 - 0.85	0.26	0.210	1.57	0.00
	Martial status (% Married)	(3-81%; M = 35%)	35	-0.10	-0.57 - 0.37	0.24	0.678	0.17	0.00
	Employment (% full-time)	(5-100%; M = 60%)	28	1.00	0.59 - 1.42	0.21	<0.001**	22.54	23.16
	Gender (% female)	(0-100%; M = 66%)	78	-0.34	-0.47 - -0.21	0.07	<0.0(**	27.03	18.27
	Risk of Bias (1-10)	(1-8; M = 5.84)	83	0.05	0.02 - 0.09	0.02	0.004 *	8.32	0.00
	Year (of publication)	(2000 - 2020)	169	0.00	-0.01 - 0.02	0.01	0.870	0.03	0.00
General	Mean age	(22-52 years; M = 35)	147	0.00	-0.01 - 0.01	0.01	0.808	0.06	0.00
	Sessions (mean)	(1-65 sessions; M = 15)	102	-0.01	-0.01 - 0	0.00	0.257	1.29	0.00
	Ethnicity (% minority)	(0-70%; M = 25%)	67	-0.47	-1.12 - 0.18	0.33	0.159	1.99	0.00
	Martial status (% Married)	(3-81%; M = 41%)	54	-0.15	-0.48 - 0.17	0.17	0.351	0.87	0.00
	Employment (% full-time)	(0-100%; M = 53%)	59	-0.10	-0.54 - 0.34	0.22	0.651	0.20	6.69
	Gender (% female)	(0-100%; M = 67%)	162	-0.31	-0.74 - 0.12	0.22	0.154	2.04	0.00
	Risk of Bias (1-10)	(1-8; M = 5.69)	169	0.05	0.01 - 0.09	0.02	0.010 *	6.60	39.01

Note. Model Outputs in Bold are significant at either * p = <.05. ** Bonferroni adjustment, p = <.00625

Discussion

The aim of this review was to provide a rigorous and comprehensive evaluation of the effectiveness of psychological therapies delivered in routine practice, and also to explore a range of potential moderators of treatment effectiveness. A broad and inclusive approach was taken, resulting in a large number of eligible studies (k = 252) and samples (k = 298) for a narrative synthesis. Of these, a large number of studies were also eligible for the meta-analysis (k = 223 [88.5%], samples = 263). This review is the largest synthesis of effectiveness studies concerning adult one-to-one psychological therapy conducted to date, expanding on prior reviews in breadth (number of studies, settings, treatment modalities) and depth (meta-analyses, risk of bias and quality appraisals, moderator analyses).

Summary of Findings

The large number of studies included in this review reflects the increase in publication of practice-based evidence; that is, the majority of studies (71.48 %) were published since 2010. Consistent with prior reviews of effectiveness, we found large pre-post treatment effects for psychological therapy in the treatment of depression, anxiety and global outcomes (i.e., psychological distress, symptoms and functioning). Method of analysis (ITT vs. completers), study continent, and methodological quality rating were significant moderators across all three treatment domains. A number of additional moderator variables were significant, but not for all domains. There was no evidence of a significant interaction between mean number of sessions and completion methodology. The finding that a large amount of PBE was conducted at stage three of the hour-glass model (add citation again for clarity) is an indication that contemporary services are largely implementing evidence-based practice.

Contribution to the Evidence Base

This review builds on prior reviews of psychological therapy effectiveness in routine care. Consistent with prior reviews, there was strong evidence that psychological

therapy leads to clinical improvements across a range of outcomes. The observed large pre-post treatment effect-size for depression outcomes (d = 0.98) was consistent with prior effectiveness reviews of depression outcomes reported by Wakefield et al. (2021) (d = 0.87) and Hans & Hiller (2013) (d = 1.13 [completers]). The large pre-post effect-size for anxiety outcomes (d = 0.98) was consistent with that reported by Wakefield et al. (2021) (d = 0.88, CI = 0.79-0.97) and the array of large effects-sizes for specific anxiety disorders reported by Stewart & Chambless (2009). Finally, the pre-post treatment effect-size for global outcomes (d = 1.01), although somewhat lower than Cahill et al. (2010) (d = 1.29) remained within the 'large' effect-size classification. This review expands on prior reviews through utilising a much larger sample of, and more diverse array of, routine services. This was also (to our knowledge) the first effectiveness review to focus on individual (i.e. one-to-one) psychological therapy.

The review found that the majority of individuals accessing psychological therapy across these studies were female. This rate of female over-representation is consistent with findings from other reviews of therapy effectiveness (e.g. 60.2% Wakefield et al., 2021) and global epidemiological studies of mental health prevalence (Seedat et al., 2009).

This review identified three moderator variables as significant across outcome domains. The finding that completer samples had consistently larger effect sizes relative to ITT samples was consistent with prior effectiveness reviews (Hans & Hiller, 2013; Wakefield et al., 2021). The consistency of this finding across a large sample of studies supports prior claims that completer samples may run the risk of providing over-inflated effect-sizes (i.e. type I error, Fergusson et al., 2002).

The finding that continent of study was a significant moderator across domains was a novel finding. Differences in therapy outcomes between continents has, to our knowledge, not been explored in prior outcome studies or meta-analyses. This review found larger effect-sizes for UK and North American studies. Caution in interpretation is required as there was high overlap in CIs. In explaining this 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 (APA2006, NICE, 2011).

The third significant moderator across domains was the continuous variable of risk of bias. Higher effect-sizes were associated with higher quality rating scores. This finding is in contrast to prior evidence which has demonstrated that greater methodological quality is associated with smaller effect sizes (e.g. Wakefield et al., 2021). The discrepancy between this finding and the extant literature is hard to explain, but may be due to differences in appraisal tools/approach to analysis. The appraisal tool in this study was designed for use with case-series designs, with a number of points based on reporting quality as opposed to methodological bias. It is possible that a tool which places greater emphasis upon other areas of bias, such as aspects of internal validity, may have provided a different pattern of results. It is also possible that treating quality rating as a sub-group moderator (i.e. as done by Wakefield et al., 2021) and not a meta-regression variable may have accounted for some of the differences in results.

There was a range of other significant moderator variables that were not significant across all three domains. Stage of the hour-glass model has not previously been explored for its potential influence on effect-size. This variable was significant, but only for the general domain. Higher effect-sizes were demonstrated for preliminary/pilot studies. It is possible that interventions within this earlier stage of development are provided with relatively more resource and impose more internal controls than established, routinely-delivered interventions within benchmarking/evaluation studies.

Therapy modality was a significant moderator for the anxiety and general domains. This finding goes against the well-established equivalence paradox in psychotherapy literature; that is, no significant difference in effectiveness between therapeutic models has consistently been shown (Wampold et al., 1997). This review found

that, for the general outcome domain, CBT produced higher effect-sizes, with no overlap in CI. Therapy modality was also significant for the anxiety domain however the overlap in CI reduces confidence in identifying a superior treatment. In explaining the superiority of CBT for global outcomes, it is possible that this was due to the inclusion of specific conditions (e.g. PTSD, OCD) for which CBT has a stronger evidence base.

An additional noteworthy finding is the differences shown in outcomes between qualified and unqualified clinicians. This was somewhat surprising as prior outcome studies have consistently found that unqualified clinicians do not produce significantly different effect-sizes to qualified clinicians (e.g. Buckley et al., 2006). Qualified therapists produced significantly higher effect-sizes for the general domain and smaller effect-sizes for the anxiety domain. A potential explanation for this is that unqualified staff are highly supervised and may therefore may be less likely to 'therapeutically drift' (Waller & Turner, 2016) from the identified therapeutic model than qualified clinicians. Training clinicians are also likely to have received more up-to-date training on evidence-based approaches and perhaps are more routinely required to engage in deliberate skills practice. It is not clear however why this set of differences would only apply to anxiety conditions, and why the reverse was shown for general outcomes.

Limitations

There are seven main limitations. First, all studies included in this review were observational by definition and design (i.e. no control group or randomisation). The absence of comparison conditions means that we are unable to rule out alternative explanations for observed effect-sizes such as regression to the mean.

Second, therapies were simply grouped by meta-therapy category. No fidelity/adherence checks were made. This means that we are unable to say with any confidence how much the interventions actually represent intended treatments.

A third limitation is the exclusive focus on self-reported outcome measures. This review of effectiveness therefore is defined by the patient only and does not necessarily extend to effectiveness of routine treatment as defined by the researcher/clinician.

Self-report measures are naturally prone to self-perception bias.

A fourth limitation concerns aspects of methodological precision. Because this review was conducted as part of a doctoral thesis the resources required to double rate all aspects of the project were not available. A substantial proportion of this large review was therefore done by a single clinician. In an effort to overcome this limitation integrity checks were conducted at several stages, with each showing substantial to near perfect reliability. A related strength is that the current review attempted to contact a large number of authors to request additional unreported information (Pearson's r) to improve precision of sample variance. It is possible however that because a large proportion of data remained unavailable, and was subsequently imputed, that this may have introduced bias. The risk of this happening was high as the response rates from authors was generally low. This was likely to have been influenced by the pragmatic decision to use a single e-mail template, with minimal tailoring. A further point on precision is that statistical interpretation of subgroup differences using confidence intervals is a somewhat conservative method; statistical differences may therefore not represent clinically meaningful differences

A fifth limitation concerns the risk of bias tool employed in this study. The tool employed was a brief measure with many of the items based on study reporting detail. This tool was selected for its perceived relevance to uncontrolled treatment studies. We would offer caution around any interpretations of risk of bias as there are many aspects of bias that this tool did not measure (e.g. fidelity, outcomes assessors).

A sixth limitation concerns the search strategy. It is highly unlikely that the search strategy used captured every available study. A 'complete' review of effectiveness research is not likely to be feasible, however 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 effectiveness research.

A final limitation is that the current review used broad outcome domains. This did not account for whether the outcome measured was of primary interest for change.

This is difficult to achieve as many studies report multiple measures, and without a specified primary measure.

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 among different 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 psychological therapy 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 included in 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. Continent, method of analysis, and risk of bias score were significant moderators across all outcome domains. 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.

Appendix

Appendix A

Table 1.11
List of search terms and limiters for systematic database search

Effectiveness Study Term	Psychological Relevence Term	Limiters
'Practice based evidence'	Psycho* OR Therap [PsycInfo]	English Language
'Routine practice'	Psycho* [CINAHL and MEDLINE]	Adult Sample
Benchmarking		
Transportability		
Transferability		
Clinical* representat		
'External valid* N0 findings		
Applicab* N0 findings		
Applicab* N0 intervention*		
'Empiric* support*' N0 treatment*		
'Empiric* support*' N0 intervention*		
'Clinical* Effective*'		
Dissem* N0 treatment*		
Dissem* N0 intervention*		
'Clinical Practice' N0 intervention*		
'Clinical Practice' N0 treatment*		
'Service deliv*' N0 intervention*		
'Service deliv*' N0 treatment*		
'Clinical* effective*' N2 evaluat*		
'Service deliv*' N0 evaluat*		
Transporting		
'Managed care setting'		
Uncontrolled		
'Community clinic'		
'Community mental health centre'		
'Clinic setting'		
'Service setting'		

Appendix B

Preference system for outcome measures

Because of the heterogeneity of outcome measures which could fit within the 'general' category the following hierarchy was used: (1) global measures of psychological distress (e.g. CORE-OM, SCL-90); (2) mono-symptomatic measures (e.g. Y-BOCS, EDE-Q). If a study used more than one measure at the same point in the hierarchy then we used the measure that had been most frequently employed in studies reviewed prior. Below is the final table of outcome measures used in the general category.

Table 1.12
Frequency of outcome measures with at least three occurances

Measure	n	Domain
BDI-II	44	Depression
BDI	34	
PHQ-9	30	
BSI (Depression)	8	
SCL (Depression)	8	
CESD-10	5	
HADS (Depression)	4	
DASS (Depression)	3	
BAI	19	Anxiety
GAD-7	19	
BSI (Anxiety)	8	
HADS (Anxiety)	7	
SCL (Anxiety)	6	
PSWQ	5	
DASS (Anxiety)	3	
CORE-OM	35	General
BSI-GSI	26	
SCL (Global)	22	
OQ-45	13	
PCL	12	
WSAS	7	
Y-BOCS	7	
EDEQ	6	
BHM	5	
GHQ	4	
CORE-10	3	
SF-36	3	

Note. Abbreviations: Beck's Depression Inventory (BDI); Patient Health Questionnaire-9 (PHQ-9); Brief Symptom Invetory (BSI); Symptom Checklist 90 Revised (SCL90R); Centre for Epidemiological Studies Depression Scale (CESD10); Depression Anxiety and Stress Scale (DASS); Hospial Anxiety & Depression Scale (HADS) Short Form-36 (SF36); Beck's Anxiety Inventory BAI); Generalised Anxeity Disorder-7 (GAD7); Penn-State Worry Questionnaire (PSWQ); CORE Outcome Measurement (CORE-OM); Outcome Questionnaire-45 (OQ45); PTSD Checklist (PCL); Work and Social Adjustment Scale (WSAS); Yale-Brown Obsessive Compulsive Scale (Y-BOCS); Eating Disorder Examination Questionnaire (EDEQ); Behavioural Health Measure (BHM); General Health Questionnaire (GHQ); Short Form-36 (SF36).

Appendix D

Table 1.13
Adapted Joanna Briggs Institute critical appraisal tool for case series designs

Criteria
1. Were there clear criteria for inclusion in the outcome study?
2. Did the outcome study have consecutive inclusion of participants?
3. Did the case series use intention-to-treat and using an appropriate method?
4. Was there clear reporting of the demographics of the participants in the study?
5. Was there clear reporting of the intervention received
6. Were the outcomes clearly reported?
7. Was there clear reporting of the presenting site(s)/clinic(s) demographic information?
8. Was statistical analysis appropriate?

Appendix E

Table 1.14
Frequency of exclusion reasons from the systematic search.

ExclusionReason	SecondaryReason	n
No effectiveness data	Aggregated outcomes measure areas	1
	Clinician rated measured	1
	No effectiveness measure	21
	No pre-post	22
	No self-report	20
	No validated measure	3
	NA	13
No individual psychotherapy	By proxy	2
	Family/couples	17
	Group therapy	67
	NA	11
No primary data	Book chapter	1
	No psychology intervention	1
	Overlap with other study	19
	Review/meta	8
	Secondary analysis	12
	NA	18
No psychotherapy	No psychology intervention	11
	NA	2
Not adult population	Not adult population	1
	Not adult Population	4
Not face-to-face	Format	2
Not naturalistic	Control group	21
	Randomisation	33
	Recruited from routine settings	2
	Stage II	4
	NA	9
Sample size	Sample size	5
No English full-text	No English full-text	1
	NA	1
No Psychotherapy	No psychology intervention	1
Original Study Used	Overlap with other study	3
OVERLAP WITH ROSEBOROUGH	Overlap with other study	2

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2020). Rmarkdown: Dynamic documents for r. https://github.com/rstudio/rmarkdown

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., Hardy, G. E., & Mellor-Clark, J. (2010). Developing and delivering practice-based evidence: A guide for the psychological therapies. Wiley-Blackwell.

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). Wiley-Blackwell. https://doi.org/10.1002/9780470687994.ch2

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck Depression InventoryII - PsycNET. APA PsycTests. https://doi.org/10.1037/t00742-000

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

Borenstein, M., Hedges, L., V, Higgins, J., P. T, & Rothstein, H., R (Eds.). (2009). 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

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed). L. Erlbaum Associates.

Cuijpers, P., Weitz, E., Cristea, I. A., & Twisk, J. (2017). Pre-post effect sizes should be avoided in meta-analyses. Epidemiology and Psychiatric Sciences, 26(4), 364–368. https://doi.org/10.1017/S2045796016000809

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

Fergusson, D., Aaron, S. D., Guyatt, G., & Hébert, P. (2002). Post-randomisation exclusions: The intention to treat principle and excluding patients from analysis. BMJ, 325(7365), 652–654. https://doi.org/

10.1136/bmj.325.7365.652

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

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

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

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

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

Jacobson, N. S., & Christensen, A. (1996). Studying the effectiveness of psychotherapy. How well can clinical trials do the job? The American Psychologist, 51(10), 1031–1039. https://doi.org/10.1037/0003-066X.51.10.1031

Kraemer, H. C., Frank, E., & Kupfer, D. J. (2006). Moderators of treatment outcomes: Clinical, research, and policy importance. JAMA, 296(10), 1286–1289. https://doi.org/10.1001/jama.296.10.1286

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. Biological Psychiatry, 59(11), 990–996. https://doi.org/10.1016/j.biopsych.2005.09.014

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 (Sixth, pp. 169–218). John Wiley & Sons, Incorporated.

Lambert, M. J., Gregersen, A. T., & Burlingame, G. M. (2004). The Outcome Questionnaire-45. In M. E. Maurish (Ed.), The use of psychological testing for treatment planning and outcomes assessment: Instruments for adults (Vol. 3, pp. 191–234). Lawrence Erlbaum Associates Publishers.

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

Li, X., Dusseldorp, E., Su, X., & Meulman, J. J. (2020). Multiple moderator meta-analysis using the R-package Meta-CART. Behavior Research Methods, 52(6), 2657–2673. https://doi.org/10.3758/s13428-020-01360-0

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

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S. J., & Kircher, J. C. (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. Journal of Consulting and Clinical Psychology, 76(1), 116–124. https://doi.org/10.1037/0022-006X.76.1.116

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. BMJ, 339, b2535. https://doi.org/10.1136/bmj.b2535

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

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

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

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

Pybis, J., Saxon, D., Hill, A., & Barkham, M. (2017). The comparative effectiveness and efficiency of cognitive behaviour therapy and generic counselling in the treatment of depression: Evidence from the 2nd UK National Audit of psychological therapies. BMC Psychiatry, 17(1), Article 215. https://doi.org/10.1186/s12888-017-1370-7

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

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., & Fonagy, P. (1996). What works for whom?: A critical review of psychotherapy research. Guilford.

Salkovskis, P. M. (1995). Demonstrating Specific Effects in Cognitive and Behavioural Therapy. In Research Foundations for Psychotherapy Practice (pp. 191–228). Wiley.

Sauer-Zavala, S., Ametaj, A. A., Wilner, J. G., Bentley, K. H., Marquez, S., Patrick, K. A., Starks, B., Shtasel, D., & Marques, L. (2019). Evaluating transdiagnostic, evidence-based mental health care in a safety-net setting serving homeless individuals. Psychotherapy, 56(1), 100–114. https://doi.org/10.1037/pst0000187

Schulz, K. F., Altman, D. G., Moher, D., & the CONSORT Group. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. BMC Medicine, 8(1), 18. https://doi.org/10.1186/1741-7015-8-18

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

Seedat, S., Scott, K. M., Angermeyer, M. C., Berglund, P., Bromet, E. J., Brugha, T. S., Demyttenaere, K., de Girolamo, G., Haro, J. M., Jin, R., Karam, E. G., Kovess-Masfety, V., Levinson, D., Medina Mora, M. E., Ono, Y., Ormel, J., Pennell, B.-E., Posada-Villa, J., Sampson, N. A., … Kessler, R. C. (2009). Cross-National Associations Between Gender and Mental Disorders in the World Health Organization World Mental Health Surveys. Archives of General Psychiatry, 66(7), 785. https://doi.org/10.1001/archgenpsychiatry.2009.36

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

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

10.1037/0003-066X.32.9.752

Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. Archives of Internal Medicine, 166(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

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

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

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empiricially, "all must have prizes.". Psychological Bulletin, 122(3), 203–215. https://doi.org/10.1037/0033-2909.122.3.203

PART II

EMPIRICAL PROJECT

2

Empirical