International Journal of Clinical and Health Psychology 24 (2024) 100446

Contents lists available at ScienceDirect

International Journal of Clinical and Health Psychology

journal homepage: www.elsevier.es/ijchp

# Benchmarks for dialectical behavioural therapy intervention in adults and adolescents with borderline personality symptoms

Julieta Azevedo [a,b,c,*], Diogo Carreiras [c,d], Caitlin Hibbs [a,b], Raquel Guiomar [c], Joshua Osborne [b], Richard Hibbs [b], Michaela Swales [a,b]

[a] School of Human and Behavioural Sciences - Bangor University, UK
[b] British Isles DBT Training, UK
[c] University of Coimbra, Center for Research in Neuropsychology and Cognitive and Behavioral Intervention (CINEICC), Portugal
[d] Miguel Torga Higher Institute, Portugal

ARTICLE INFO

*Keywords:*
Benchmarking
DBT
Borderline symptoms
BPD
EQ-5D
BSL
DERS

ABSTRACT

*Background:* Dialectical Behaviour Therapy (DBT) is a multi-component cognitive behavioural intervention with proven efficacy in treating people with borderline personality disorder symptoms. Establishing benchmarks for DBT intervention with both adults and adolescents is essential for bridging the gap between research and clinical practice, improving teams' performance and procedures.

*Aim:* This study aimed to establish benchmarks for DBT using the EQ-5D, Borderline Symptoms List (BSL) and Difficulties in Emotion Regulation Scale (DERS) for adults and adolescents.

*Methods:* After searching four databases for randomised controlled trials and effectiveness studies that applied standard DBT to people with borderline symptoms, a total of 589 studies were included (after duplicates' removal), of which 16 met our inclusion criteria. A meta-analysis and respective effect-size pooling calculations (Hedges-g) were undertaken, and heterogeneity between studies was assessed with $I^2$ and Q tests. Benchmarks were calculated using pre–post treatment means of the studies through aggregation of adjusted effect sizes and critical values.

*Results:* DBT aggregated effect sizes per subsample derived from RCTs and effectiveness studies are presented, along with critical values, categorised by age group (adults vs adolescents), mode of DBT treatment (full-programme vs skills-training) and per outcome measure (EQ-5D, BSL and DERS).

*Conclusions:* Practitioners from routine clinical practice delivering DBT and researchers can now use these benchmarks to evaluate their teams' performance according to their clients' outcomes, using the EQ-5D, BSL and DERS. Through benchmarking, teams can reflect on their teams' efficiency and determine if their delivery needs adjustment or if it is up to the standards of current empirical studies.

## Introduction

Dialectical behaviour therapy (DBT) is a multi-component and integrative treatment that synthesises behavioural theory, principles of Zen practice and dialectics (Linehan, 1993). This therapy was initially designed to treat people with chronic suicidality who had received a diagnosis of Borderline Personality Disorder (BPD). It was the first treatment to consistently demonstrate clinical efficacy with this client group and continues to be the most supported (Stoffers-Winterling et al., 2012, 2022).

DBT comprehensive treatment, also referred to as full-programme, includes four modes of treatment: individual psychotherapy, skills training, phone coaching, and consultation for therapists (Linehan, 2015; Swales & Heard, 2017). The skills are taught in a group format (and strengthened in individual therapy) and include modules on emotional regulation, mindfulness, distress tolerance and interpersonal effectiveness. For less severe clients (without suicide ideation or active self-harm), non-comprehensive skills training can be offered (which ideally should still include a consultation team and phone coaching) (Linehan, 2015; Valentine et al., 2015).

BPD is a severe disorder characterised by emotional, interpersonal and intrapersonal instability, impulsivity, self-harm behaviours and

* Corresponding author at: biDBT Training, Wrexham Technology Park, Croesnewydd Hall, Wrexham LL13 7YP, UK.
*E-mail address:* julietazevedo@gmail.com (J. Azevedo).

https://doi.org/10.1016/j.ijchp.2024.100446
Received 21 September 2023; Accepted 27 January 2024
Available online 6 February 2024

feelings of emptiness and abandonment (American Psychiatric Association, 2022; Biskin, 2015; D'Aurizio et al., 2023; Zanarini et al., 2011). Borderline symptoms usually begin in adolescence and are associated with negative affect and a high risk of suicide, which remains as high as 10 % over a 27-year course (APA, 2022; Videler et al., 2019). This suicide rate is partially due to emotional dysregulation (Mirkovic et al., 2021) and low levels of quality of life (IsHak et al., 2013). When assessing the efficacy of DBT, researchers have focused on decreasing borderline symptomatology, increasing skills use (e.g. emotion regulation), as well as decreasing suicide and self-harm behaviour (Stoffers--Winterling et al., 2022) and quality of life interfering behaviours, thus looking for improvements in quality of life (Carter et al., 2010; Chakhssi et al., 2021; van Asselt et al., 2009).

Scientists, healthcare professionals, and decision-makers worldwide recognise that evidence-based clinical interventions do not translate easily into routine practice due to numerous implementation challenges (Eldh et al., 2017), resulting in a research-practice gap (Chorpita & Daleiden, 2014; Gotham, 2006; Teachman et al., 2012). One recognised barrier is the practitioner's scepticism (Lilienfeld et al., 2013), partially relating to concerns about whether clients and therapists in RCTs are representative of those in typical clinical settings (Hunsley & Lee, 2007).

Delgadillo et al. (2014) conducted a comprehensive examination of the widespread implementation of empirically supported treatments in routine practice and assessing outcomes in everyday clinical settings. In doing so, they identified notable challenges associated with the measurement, definition, and comparative analysis of clinical outcomes. For future benchmarking initiatives, the authors emphasised the importance of comparing effect size estimates derived from condition-specific measures with those obtained using more generalised distress measures. Furthermore, the authors emphasised the need for a nuanced approach to benchmarks, considering contextual, diagnostic, and population factors in specific settings.

Benchmarking is an outcome assessment strategy used to assess clinical services (Eisen & Dickey, 1996) by evaluating patients' data and clinical outcomes (Lovaglio, 2012). In clinical psychology, a benchmark represents a standard of care or best practice for specific treatments or interventions, working as a reference point for comparison or measurement. The key takeaway is that benchmarking fundamentally aims to enhance practices and processes, ultimately improving outcomes through establishing care standards, identifying and delivering effective treatments, and monitoring care quality (Lloyd, 2004).

Only one study, to our knowledge, has established benchmarks for DBT treatment in adult outpatients with BPD (Washburn et al., 2018). However, this study had strict inclusion criteria and focused exclusively on RCT studies delivering DBT full-programme measuring depression, anger and self-harm. Moreover, Washburn et al. (2018) aggregated studies that used different instruments to measure the above-mentioned outcomes, which has limitations when drawing benchmarks since they vary in their specificity and reactivity (Minami et al., 2008). Additionally, considering their database search was done in 2016, a more up-to-date review is required, especially one including DBT skills groups intervention and including adolescents. Finally, for a comprehensive team performance evaluation, it is crucial to employ measures assessing health-related quality of life to inform cost-effectiveness and mechanisms of change within the treatment.

The project was developed as part of an ongoing collaboration between Bangor University, British Isles DBT Training, and NHS England, formerly Health Education England (HEE), to significantly increase the number of trained DBT clinicians embedded in active DBT programmes. As part of the project, there was an aim to benchmark the clinical outcomes of teams as a means to assess the effectiveness of the training programme in producing a return on the training investment.

The first challenge faced was the lack of a body of evidence to draw from. Due to a general lack of previous studies benchmarking the constructs we were interested in assessing, there was a need to first establish benchmarks based on empirical studies, including both effectiveness and RCT studies, to obtain a suite of benchmarks for teams to use to compare against their own routine clinical practice outcomes.

The current study aims to provide benchmarks for DBT treatment by meta-analysing data from quality RCTs and effectiveness studies delivering standard DBT (full-programme/ or stand-alone skills training) to adults and adolescents with BPD symptoms and that used the same instruments to measure emotion regulation, health-related quality of life and borderline symptoms.

## Methods

### *Benchmarking methodology*

In terms of the steps to start benchmarking, we considered the plan described by Lloyd (2004) and the considerations of Bayney (2005), beginning with the selection of the data collection method and establishing the measures to act as benchmarks. Furthermore, we reviewed specific studies that benchmarked mental health interventions, adopting their recommendations and structure (Delgadillo et al., 2014; Minami et al., 2007; Weersing & Weisz, 2002). Minami et al.'s (2008) proposal to withdraw benchmarks from clinical trials was followed, given its clarity and high citation rate in other studies. In order to review the published clinical trials, a thorough database search was conducted, followed by a meta-analysis of the collected data. This analysis included published studies using three selected measures (described below) to measure DBT intervention outcomes in adults and adolescents with borderline features so that DBT teams can use them to assess their performance. Weersing and Weisz (2002) advise estimating benchmarks based on studies that used an intention-to-treat (ITT) approach; however, excluding studies without clear ITT information would have led to a substantial loss of data, mainly from effectiveness studies, hindering the establishment of robust and representative benchmarks. Hence, we decided to include studies that described only completers data, reporting the used method for all the studies.

### *Selection of outcome measures*

The considerations of Delgadillo et al. (2014) on selecting outcome measures to benchmark were followed: weighing up the research evidence of the tools, sensitivity to change with the intended population, ease of administration and interpretation and measures that are free to use and easily accessible.

Minami et al. (2008) indicated that ideally, the selection of clinical trials for inclusion for a benchmark should utilise identical measures to ensure they match in specificity and reactivity. Following this guideline, we selected three measures to assess: health-related quality of life, which in addition allows for clinical and economic appraisal; difficulties in regulating emotions, which is considered an important mechanism of change in DBT; and a disorder-specific measure for borderline symptoms.

### *Selected instruments and rational*

The *EQ-5D-3 L* (EuroQol Research Foundation, 2018), henceforth referred to as "EQ-5D" is a widely used generic measure of health status consisting of two parts. The first part (the descriptive system) assesses health in five dimensions (mobility, self-care, usual activities, pain/-discomfort, anxiety/depression), each of which has three levels of response (no problems, some problems, extreme problems/unable to), providing a health state profile. Each health state is assigned a summary index score based on societal preference weights for the health state. These weights, or utilities, are used to compute Quality-Adjusted Life Years (QALYs) in health economic analyses. Health state index for the United Kingdom ranged from 1 ('perfect' health state), with higher scores indicating higher health utility, to −0.543. EQ-5D's negative range in the interval corresponds to health status 'worse' than death, which has face validity in suicidal populations and can vary slightly

between countries.[1]

The second part of the questionnaire consists of a Visual Analogue Scale (VAS) on which patients rate their perceived health from 0 (the worst imaginable health) to 100 (the best imaginable health). EQ-5D-3 L has revealed good test-retest reliability across studies ranging from 0.53 to 0.83 (Buchholz et al., 2018). The EQ-5D questionnaire is cognitively undemanding, taking only a few minutes to complete. A version for adolescents is also available (EQ-5D-3L-Y), with the same components and some adjustments to the language only. The anchor points for the EQ-5D, with 0 representing death and less than 1 a state worse than death, make the EQ-5D a perfect measure conceptually for treating a client group where a central feature is the desire to die by suicide and in whom suicide rates are high. The measure also aligns well with DBT's primary aim to create a life worth living (i.e., utility score is sensitive to changes in psychotherapy for people with BPD (van Asselt et al., 2009).

The *Difficulties in Emotion Regulation Scale* (DERS; Gratz & Roemer, 2004) measures difficulties in regulating emotions, which is considered one of the core problems experienced by people with a diagnosis of BPD (Glenn & Klonsky, 2009). This instrument is one of the most widely used to measure this construct (Sloan et al., 2017), and it has been used effectively to detect changes in interventions for BPD (Stoffers-Winterling et al., 2022). This self-report questionnaire assesses six components of difficulties in regulating emotions: nonacceptance of emotional responses, difficulty engaging in goal-directed behaviour, impulse control difficulties, lack of emotional awareness, limited access to emotion regulation strategies, and lack of emotional clarity. The 36 items are rated on a 5-point Likert scale ranging between 1 ("almost never") and 5 ("almost always"). The DERS has shown good psychometric properties, with high internal consistency, a Cronbach's α of 0.93, and test-retest reliability of 0.88. There is a short form with 18 items (DERS-SF; Kaufman et al., 2016), and both versions are validated to use with adults and adolescents (DERS-36, Gratz & Roemer, 2004; Neumann et al., 2009; DERS-SF, Weinberg & Klonsky, 2009).

The *Borderline Symptoms List* (BSL; Bohus et al., 2009) is a self-report questionnaire with 23 items, which aims to assess BPD symptoms. This scale's items cover BPD diagnostic criteria (e.g., affective instability, recurrent suicidal behaviour or threats, non-suicidal self-injury, and transient dissociative symptoms) and borderline-typical processes such as self-criticism, emotional vulnerability, self-disgust and loneliness. Items are rated on a 5-point Likert scale from 0 ("not at all") to 4 ("very strong"). This instrument was chosen because it has recently been updated in line with the criteria in DSM-5 and was developed based on the experiences of both clinical experts and input from people with a diagnosis of BPD (Kleindienst et al., 2020). The BSL-23 has a single-factor structure and excellent psychometric properties, with high internal consistency, a Cronbach's $\alpha$ of 0.97, and test-retest reliability of 0.82 (Bohus et al., 2009). Moreover, it is an instrument tested on over 1000 adults with defined cut-off scores and severity levels (none or low, mild, moderate, high, very high, and extremely high) that facilitates score interpretation. In their study to propose severity levels, Kleindienst et al. (2020) specify that a score of 1.5 is typical for someone with a diagnosis of BPD but highly extreme compared to someone without a psychiatric diagnosis.

While many DBT outcome studies focus on suicidal behaviour and Non-Suicidal Self-Injury (NSSI), we decided against endeavouring to benchmark these outcomes for several reasons. First, studies use widely varying measures of these constructs, making comparison difficult. Second, the most reliable and valid measures of this construct used in the literature are typically interview-based and beyond the capacity of many routine clinical services. Finally, not all clients within a DBT service will necessarily be suicidal or engaging in NSSI. These reasons make measures of these constructs less applicable to all clients in a service. Additionally, self-harm benchmarks are already available in Washburn et al.'s (2018) study.

[1] For detailed information per country consult - https://euroqol.org/eq-5d-instruments/eq-5d-3l-about/population-norms/

### *Search procedures*

Our research team performed a literature search in September 2022 on the efficacy of DBT for the treatment of BPD. The search terms used were "dialectical behavior therapy" or equivalent (e.g., "DBT", "dialectical behaviour therapy"), "borderline personality disorder" or related (e.g., "BPD", "borderline symptoms"), and "randomised controlled trial", "systematic review" or similar (e.g., "RCT", "meta-analysis", "clinical trial"). The search was conducted in the following databases: PubMed/MEDLINE, Embase, PsycINFO and Web of Science. The study selection was then performed in two phases. Firstly, through screening the titles and abstracts, and secondly, the full texts, according to the inclusion and exclusion criteria. We also searched for systematic reviews or meta-analyses on DBT efficacy or effectiveness to ensure that we had included all relevant studies. The search was repeated in May 2023 to check if new studies had been published, but no additional articles met our inclusion criteria.

We included only original articles published in journals with peer review and used the PICO (Population, Intervention, Comparison, Output) approach, which is depicted in Table 1. To include the articles in our review the following characteristics were also confirmed:

(a) RCT or effectiveness study design
(b) Conducted in public mental health and community outpatient settings
(c) Written in English, French, Spanish, Portuguese or German
(d) Full text available (either in open access or through subscription - database search from Bangor University and University of Coimbra)
(e) Good quality of the articles using the JBI critical appraisal tools (Joanna Briggs Institute, 2017)

A similar approach was used for RCTs and effectiveness studies, so comparison or control groups were not accounted for comparison. It is also important to clarify that the studies selected reported that participants engaged in pharmacotherapy and maintained the treatment provided by their psychiatrist or GP (as usual in most psychotherapy-focused studies). Participants were not required to remain on consistent medication; therefore, medication changes may have occurred.

A flow diagram with the identification and selection process is depicted in Fig. 1. A total of 1253 studies were identified and uploaded into Rayyan (https://rayyan.ai/), a free web platform that helps expedite the initial screening of abstracts and titles for systematic reviews (Ouzzani et al., 2016) and 589 duplicates were eliminated. We included only original articles published in journals with peer review.

A total of 627 studies were excluded after applying our criteria. We selected 37 studies for full-text detailed reading and quality assessment. Of these, 21 were excluded for the following reasons: sample already included in one of the selected studies (5 studies); modified DBT (for example, DBT Prolonged Exposure); different length or adaptation for non-BPD populations (5 studies); necessary data not reported, namely the outcomes *M* and *SD* (5 studies); low study quality (according to JBI critical appraisal tools) (2 studies); mixed samples (4 studies).

The main features (e.g., study design, sample size) of the final 16 selected studies can be found in Table 2, of which 13 used adult samples and three adolescent samples. These studies were conducted in eight different countries: USA (4), Ireland (3), Canada (3), Australia (2), UK (1), Germany (1), Netherlands (1) and Norway (1).

### *Assessment of studies' quality*

The full text of 37 studies was read thoroughly, and they were then assessed in terms of methodological quality and quality of the report of

**Table 1**
PICO approach with inclusion and exclusion criteria applied.

| PICO | Inclusion Criteria | Exclusion Criteria |
|---|---|---|
| P - Population | • Adolescents from 12 to 18 years of age with borderline symptoms, and adults with ages ranging from 18 to 65 with BPD diagnosis. | • Older adults<br>• PTSD, patients selected based on BPD+other diagnosis. |
| I - Intervention | • Dialectical Behaviour Therapy (full programme) 6 to 12 months<br>• DBT Skills training only, with a minimum duration of 16 weeks to 24, using manualised treatment. | • Adapted interventions with different treatment length or frequency of sessions<br>• Skills interventions without the four skills modules: emotion regulation, distress tolerance, interpersonal effectiveness, mindfulness. |
| C - Comparison | • Practitioners had to have been adequately trained in DBT.<br>• Within-group comparisons between baseline and post-treatment scores<br>• Focus on the intervention received and not the comparison group. | • Concomitant psychotherapy other than DBT or enhanced treatment<br>• Different assessment timepoints, preventing comparison at baseline and post treatment |
| O - Outcomes | • Health-related quality of life - EQ-5D difficulties in emotion regulation - DERS, and borderline symptoms list - BSL-23 | • Studies which did not report pre-post M/SD from EQ-5D, DERS, or BSL. |

*Note: M* = Mean; *SD* = Standard Deviation; BPD = Borderline Personality Disorder; PTSD = Post-Traumatic Stress Disorder; DBT = Dialectical Behaviour Therapy; DERS = Difficulties in Emotion Regulation Scale; BSL = Borderline Symptoms List.

data, as well as study bias with the JBI critical appraisal tools (Joanna Briggs Institute, 2017). The checklist for randomised controlled trials (13 items) was used for RCTs, and the quasi-experimental studies checklist (non-randomised experimental studies; nine items) was used for the effectiveness studies. The authors of the JBI tool do not provide a cut-off for the scale, stating the result depends on the reason for using the scale and how rigorous a user of the scale would like to be with the accepting/excluding criteria. We decided to accept RCTs above eight (from a maximum possible score of 13), considering that some items result from methodological choices or blinding related parameters that were not relevant considering the benchmark goal for routine practice.

The checklist for quasi-experimental studies includes four items related to having a comparison group, which depresses the score to a maximum of five. Thus, provided the studies met the five quality criteria, we included effectiveness studies without a control group.

Reliability was ensured by having two independent assessors who rated the quality of the studies separately. Any discrepancies identified were resolved through consensus discussions with all assessors, resulting in the final presented score (see Table 2).

### *Data extraction process and calculation of benchmarks*

The pre–post *M* and *SD* for DERS, BSL and EQ-5D, as well as the sample size of the groups that received DBT, were retrieved from the papers. In regards to the EQ-5D, we found three studies which only reported the EQ-5D VAS scores and not the utility scores. Therefore, it was decided to contact the authors to request their datasets and their permission to use their outcomes to generate benchmarks. The authors Sinnaeve et al. (2018) and McMain et al. (2009, 2022) granted permission to use their data and provided the necessary information and databases at the beginning of 2023. Once all the necessary information from the studies was gathered, an analysis for heterogeneity was performed, followed by aggregated benchmarks for different data groupings. We found different lengths for full-programme: some studies delivered the full-programme for 12 months, and some for 6 months. A mean comparison was performed to search for significant differences (with respect to the instruments we were interested in) for these two programme lengths, and none were found in the outcomes of interest. A recent McMain et al. (2022) study compared 6 and 12-month programmes and while differences were found in suicidality across time and hospital admissions, no differences were observed in pre–and post-outcomes regarding the variables of interest in our study. For that reason, we decided to group all the studies that delivered full-programme (6 M and 12 M), taking into consideration that the 6-month has the same content as the 12-month programme (the content of the programme is delivered across six months, and then it is repeated in the following six to consolidate treatment gains). The studies were then grouped according to the type of trial, the modes of DBT delivered, the instrument they used and their age group.

Heterogeneity between studies was assessed with standard meta-analysis statistical methods. The percentage of effect size variability was evaluated using the $I^2$ formula (values of 0 %, 25 %, 50 %, and 75 % indicate no observed low, moderate, or large degrees of heterogeneity, respectively). The Cochran's *Q*-test was used to determine whether statistical heterogeneity existed. To explore possible sources of heterogeneity, subgroup analyses were performed according to the type of treatment (full-programme vs skills programme), age (adolescent vs adult) and trial (effectiveness vs RCT). To obtain pooled estimates and 95 % confidence intervals (CI) for the mean difference (*MD*) between post-treatment and baseline scores in each subgroup, a meta-analysis was performed using Hedges' *g* method (Hedges, 1981). Given that significant heterogeneity was found, random-effects models (Higgins et al., 2009) were used with the restricted maximum-likelihood method (DerSimonian & Laird, 1986). The inverse variance weighting approach was applied to assign weights to studies, giving less weight to smaller studies than larger ones. Forest plots were used to represent the pooled

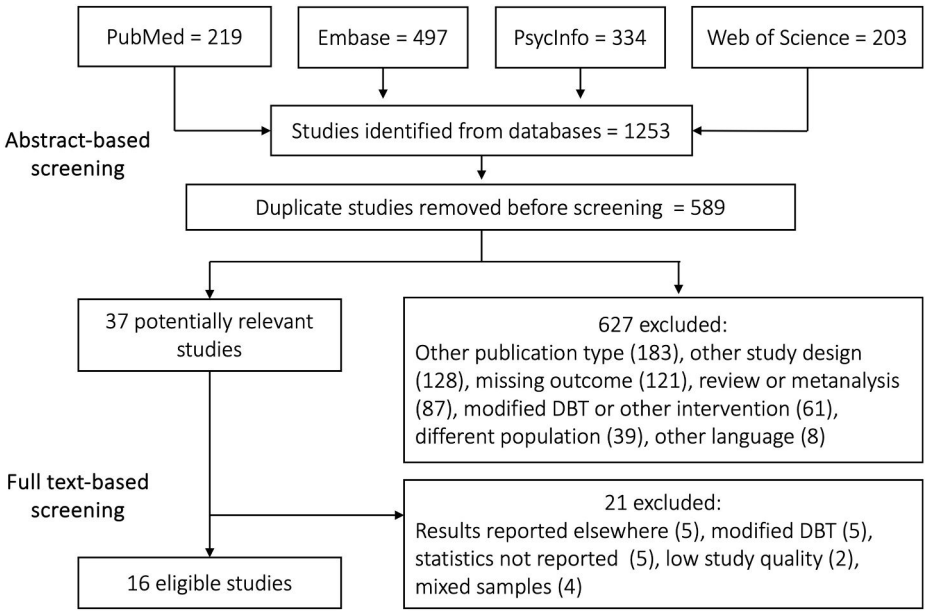


**Fig. 1.** Diagram flow of selected studies.

**Table 2**
Main features of selected studies with Dialectical Behavioural Therapy (DBT) intervention for adults and adolescents ($N = 16$).

| Authors | Study design | DBT mode | Age groups | DBT Length (in months) | Sample size at baseline | Gender % *(F, M)* | Dropout rate% | ITT | Outcome measures | JBI appraisal |
|---|---|---|---|---|---|---|---|---|---|---|
| Walton et al. (2020) | RCT | Full-programme | Adults | 12 | 81 | 77, 23 | 35 | Yes | DERS | 10/13 |
| Stiglmayr et al. (2014) | Effectiveness | Full-programme | Adults | 12 | 70 | 84, 16 | 40 | Yes | BSL | 5/9 |
| Lyng et al. (2019) | Effectiveness | Full-programme | Adults | 12 | 37* | 78, 22 | 21 | No | BSL | 8/9 |
| Kuehn et al. (2020) | RCT | Full-programme | Adults | 12 | 66* | 100, 0 | 20 | Yes | DERS | 10/13 |
| Barnicot & Crawford (2018) | Effectiveness | Full-programme | Adults | 12 | 58 | 72, 28 | 47 | Yes | DERS | 7/9 |
| Goodman et al. (2014) | Effectiveness | Full-programme | Adults | 12 | 11 | 82, 18 | 23 | No | DERS | 6/9 |
| Kells et al. (2020) | Effectiveness | Full-programme | Adults | 6 | 100 | 71, 29 | 49 | No | DERS | 5/9 |
| McMain et al. (2017) | RCT | Skills | Adults | 5 | 42 | 83, 17 | 31 | Yes | BSL, DERS | 9/13 |
| McMain et al. (2009) | RCT | Full-programme | Adults | 12 | 90 | 90, 10 | 38 | Yes | EQ-5D | 11/13 |
| McMain et al. (2022) | RCT | Full-programme | Adults | 12 | 240* | 79, 21 | 30 | Yes | BSL, EQ-5D | 10/13 |
| Heerebrand et al. (2021) | Effectiveness | Skills | Adults | 5 | 114 | 92, 8 | 27 | No | BSL | 7/9 |
| Rizvi et al. (2017) | Effectiveness | Full-programme | Adults | 6 | 50 | 80, 20 | 32 | Yes | BSL, DERS | 8/9 |
| Sinnaeve et al. (2018) | RCT | Full-programme | Adults | 12 | 42 | 95, 5 | 37 | Yes | EQ-5D | 8/13 |
| Mehlum et al. (2014) | RCT | Full-programme | Adolescents | 5 | 39 | 87, 13 | 26 | Yes | BSL | 11/13 |
| Gillespie et al. (2019) | Effectiveness | Full-programme | Adolescents | 6 | 152* | 85, 15 | 22 | No | BSL | 6/9 |
| Berk et al. (2018) | Effectiveness | Full-programme | Adolescents | 6 | 24 | 92, 8 | 8 | Yes | DERS | 7/9 |

*Note.* ITT = intention-to-treat; JBI = Joanna Briggs Institute.
* Studies with two different samples (experimental + control) that received DBT intervention.

estimates visually within subgroups and across studies. The meta-analyses of subgroups were conducted using the *meta* package (Balduzzi et al., 2019) in R statistical software version 4.1.0 (R Core Team, 2017). The statistical significance threshold was set at 0.05.

A two-step process was followed: first, we aggregated the pre-treatment–post-treatment data (means = *M*, and standard deviations = *SD*) within the studies which used the same outcome instrument, to calculate a single pre-treatment–post-treatment effect size estimate (*d*+).

The second step involved aggregating each effect size estimate to

obtain a single pre-treatment–post-treatment effect benchmark $d + j$ for each outcome measure category (see formulas, Minami et al., 2008). Additionally, in order to establish that an effect size estimate obtained from clinical settings is equivalent to efficacy benchmarks, previous literature suggested using a critical value which is dependent on the sample size of the clinical setting data. We adopted the minimum effect size of $d_{min} = 0.2$ as the criterion for clinically significant differences between benchmarks and the treatment effect size estimates for a range-null hypothesis test, (Minami et al., 2008) and we reported them only for the aggregated studies (RCT + Effectiveness).

A total of 673 adult participants were included in these benchmark calculations, 534 receiving DBT full-programme and 139 receiving Skills group intervention. A total of 173 adolescents receiving full-programme were also included.

## Results

As the $Q$ and $I^2$ statistics for homogeneity indicated that effect size estimates were heterogeneous, the reported benchmarks should not be considered an estimate of a single population parameter but rather the mean of the effect sizes estimates (Shadish & Haddock, 1994).

Results of the subgroup meta-analyses indicate that DBT significantly improved both difficulties in emotion regulation and BPD symptoms (see Figs. 2–4), for the full-programme ($MD$ = - 34.62, $k$ = 7, 95 % CI [−41.71, −27.53]; $MD$ = −0.76, $k$ = 9, 95 % CI [−1.02, −0.49], respectively), versus skills modes ($MD$ = −31.76, $k$ = 2, 95 % CI [- 46.35, −17.16]; MD = −0.81, $k$ = 2, 95 % CI [−1.15, −0.48], respectively), and effectiveness studies (MD = −30.84, $k$ = 5, 95 % CI [−38.37, −23.32]; MD = −0.77, $k$ = 7, 95 % CI [−1.13, −0.41], respectively) vs RCTs ($MD$ = −37.69, $k$ = 4, 95 % CI [−47.62, −27.76]; $MD$ = −0.73, $k$ = 4, 95 % CI [−0.86, −0.60], respectively). When analysing adolescents and adults separately, subgroup analyses indicate that DBT was effective for both adolescents and adults, as measured by BSL ($MD$ = −0.72, $k$ = 3, 95 % CI [−0.91, −0.54]; $MD$ = - 0.78, $k$ = 8, 95 % CI [−1.08, −0.47], respectively). For the DERS, only one study included adolescents, so it is only possible to pool the adults' results, which were also significant ($MD$ = −34.43, $k$ = 8, 95 % CI [−40.96, −27.90]). For EQ-5D (USc and VAS), subgroup analyses were not performed, given that all the studies found were RCTs, with adults, and delivered full-programme. The meta-analyses showed that DBT full-programme applied to adults in an RCT design significantly improves both EQ-5D utility scores (EQ-5D US) and perceived health (EQ-5D VAS; $MD$ = −0.06, $k$ = 4, 95 % CI [0.02, 0.09]; $MD$ = 7.31, $k$ = 4, 95 % CI [3.52, 11.11], respectively).

### *Benchmarks for adults*

The weighted pre and post-treatment (full-programme) $M$ and $SD$ of the RCTs and effectiveness studies were aggregated per assessment measure (BSL, DERS and EQ-5D) and per type of study (RCTs and Effectiveness studies), as shown in Table 3.

The aggregated treatment efficacy benchmarks for DBT full-programme's intervention are displayed in Table 4, grouped by instrument and type of trial, and for DBT skills intervention in Table 5 (only RCTs, because there were no effectiveness studies), showing the overall aggregation of the means and effect sizes of the studies, as well as the calculated critical value when possible. Additionally, in Figs. 5 and 6, it is possible to see the critical values and effect sizes by trial type for BSL and DERS, according to the sample size estimations. For studies reporting skills-only intervention, only the BSL and DERS studies produced benchmarks (in adults) because there were no studies using the EQ-5D. There were no significant mean differences either in pre- or post-outcomes between the effectiveness studies and the RCTs.

### *Benchmarks for adolescents*

Only three studies with adolescents were possible to include to establish benchmarks, and those studies used either BSL ($n$ = 2) or DERS ($n$ = 1), with no studies using EQ-5D. From these studies, we decided to include both samples of adolescents (experimental and control group) of Gillespie et al. (2019) because they differ only in the treatment length (16 vs 24 weeks), with no significant differences between groups in the studies using the BSL. The four data samples retrieved from the three studies with adolescents are presented in Table 6, aggregated per instrument when possible. The aggregated effect sizes for BSL are presented in Table 7. We also calculated if from Berk et al. (2018) study, which used the DERS and revealed a large effect size ($d$ = 1.095).

## Discussion

The interest in benchmarking psychological interventions is on the rise. This seems to be driven by policy changes and requests from regular service providers (Delgadillo et al., 2014; Moroz et al., 2020). As a result, there is an increased emphasis on recognising the significance of setting global standards for mental health systems. Providing benchmarks for empirically supported treatments is essential to serve as a reference point for national implementation efforts. Currently, the main challenges when benchmarking any treatments are establishing common measures to benchmark against and the lack of consensus on what and how to benchmark. This study therefore offers a proposal for common ground, providing benchmarks to assess standard DBT, based on RCTs and effectiveness studies using our three selected instruments (EQ-5D-3 L, BSL and DERS). This will allow clinical services that are using these instruments to compare their performance against these standards. In mental health treatments, benchmarking can play an essential role in evaluating the effectiveness of the treatments being offered to ensure outcomes are being compared to standards.

Even though DBT is now a widely used treatment for people with a diagnosis of BPD, with substantial evidence of its efficacy (Gillespie et al., 2022; Stoffers-Winterling et al., 2022), when performing a broad literature search through four widely used databases, it became clear that a multiplicity of measures are in use with little common ground. After applying our inclusion criteria that aimed to find robust studies using our chosen measures (while delivering DBT full-programme or skills), we were only able to select 16 studies from the 664 studies identified. The small number of studies confirms the difficulties previously mentioned by other authors, making it hard to compare their outcomes against the literature (Delgadillo et al., 2014).

The retrieved studies served as the basis for aggregating outcomes for the BSL, DERS, and EQ-5D, enabling us to establish benchmarks for both the full DBT programme and the skills programme for adults, as well as the full-programme for adolescents (using the BSL alone). Meta-analyses were performed per instrument and per subgroup (by type of trial, age group and DBT mode). Results indicated improvements in emotion regulation, a decrease in BPD symptoms, and increased health-related quality of life (EQ-5D). On the one hand, the significant heterogeneity observed across most studies (except BSL in adolescents) emphasises the need for caution when interpreting and generalising these calculated benchmarks. On the other hand, the subgroup analysis proves valuable as it reveals that DBT leads to significant improvements in the assessed instruments across various contexts (consider the controlled and rigorous nature of RCTs vs effectiveness studies), modes (comprehensive DBT vs skills only), and samples (adolescents vs adults). This implies that the studies used to derive these benchmarks are robust sources of data. They provide substantial support for the efficacy of DBT interventions and validate the sensitivity of the chosen instruments to changes within this population and treatment context.

Until this point, researchers and clinicians applying DBT intervention with our selected instruments were limited to comparing their findings with existing studies in isolation, without a clear understanding of how representative these studies were of the overall empirical data. Our article provides aggregated metrics, allowing for an examination of the means, standard deviations and effect sizes of RCTs and effectiveness

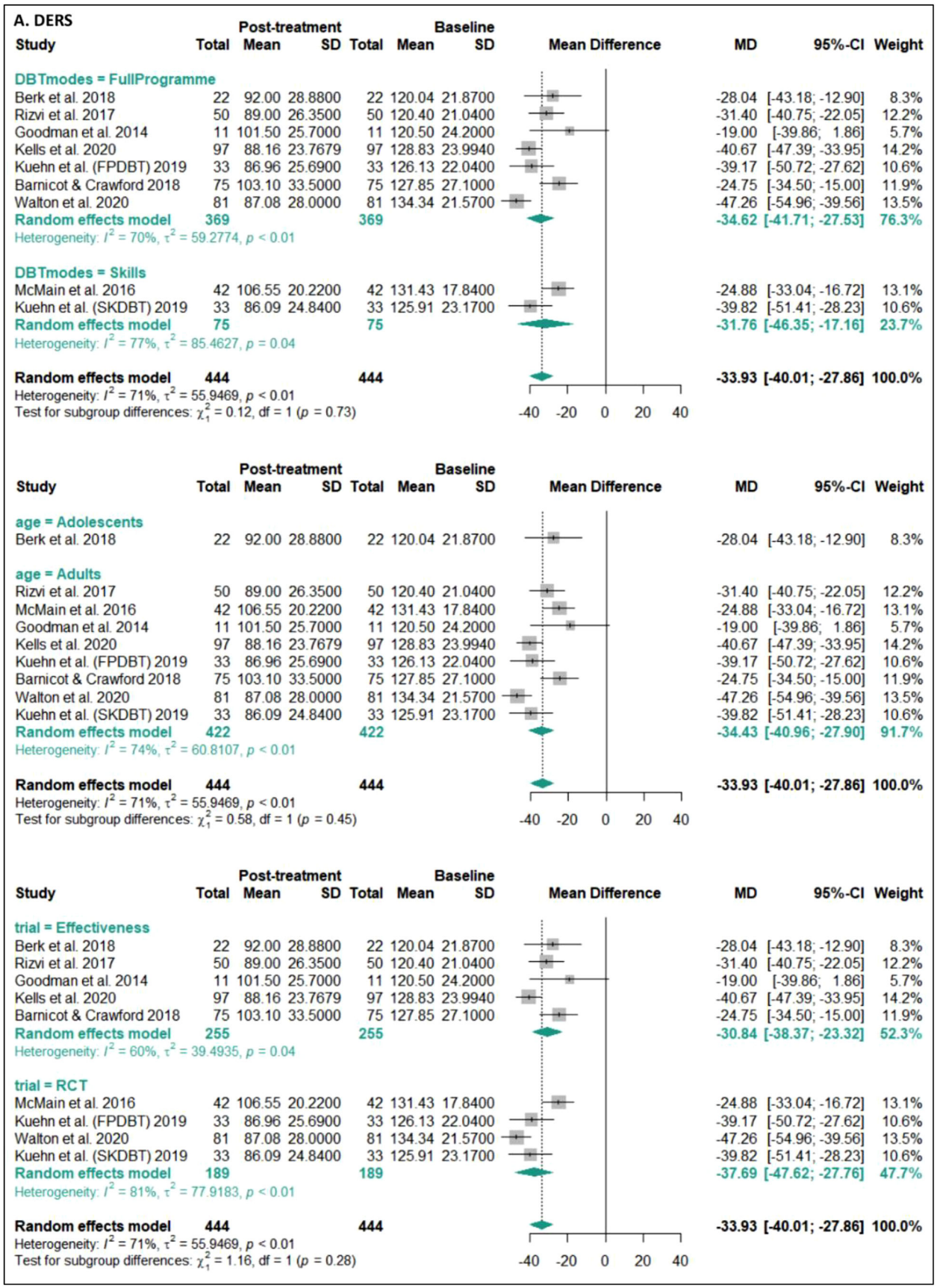


**Fig. 2.** Forest-plots of the subgroup meta-analyses with DERS assessment tool.

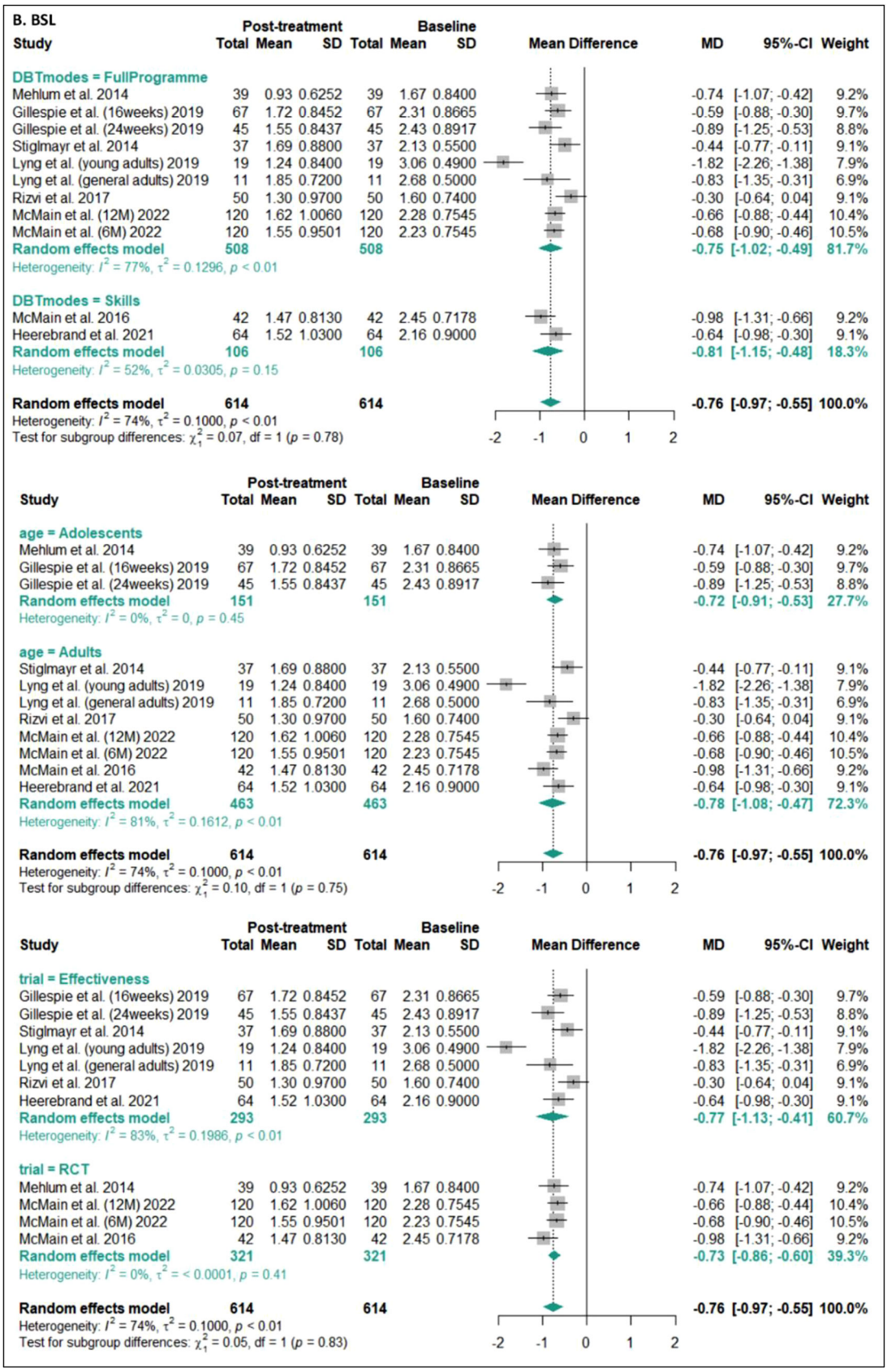


**Fig. 3.** Forest-plots of the subgroup meta-analyses with BSL assessment tool.

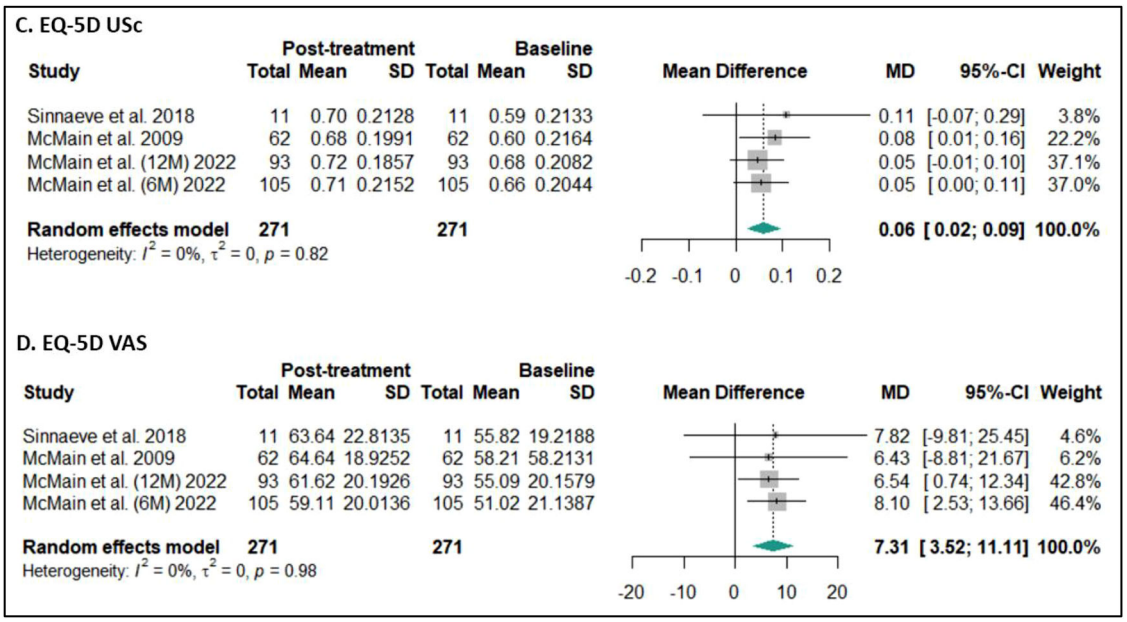


**Fig. 4.** Forest-plots of the meta-analyses with EQ-5D-3 L assessment tool Utility Score (USc) and Visual Analogue Scale (VAS).

**Table 3**
Pre- and post-treatment outcomes for adult samples that received DBT treatment in the selected studies (full-programme).

| | DBT details | n | Pre-treatment *M* ± *SD* | Pre-treatment *interval* | Post-treatment *M* ± *SD* | Post-treatment *interval* |
|---|---|---|---|---|---|---|
| **Borderline Symptoms List (BSL)** | | | | | | |
| Aggregated RCTs | 6–12 months | 240 | 2.25 ± 0.75 | [1.50 - 3.00] | 1.58 ± 0.98 | [0.6 - 2.56] |
| Aggregated Effectiveness | 6–12 months | 67 | 2.11 ± 0.82 | [1.28 – 2.93] | 1.46 ± 0.93 | [0.53 - 2.39] |
| **Difficulties in Emotion Regulation Scale (DERS)** | | | | | | |
| Aggregated RCTs | 6–12 months | 114 | 3.62 ± 0.61 | [3.01–4.23] | 2.42 ± 0.75 | [1.67 - 3.17] |
| Aggregated Effectiveness | 6–12 months | 233 | 3.48 ± 0.67 | [2.81–4.15] | 2.61 ± 0.76 | [1.89 - 3.41] |
| **EQ-5D (Utility Scores and VAS)** | | | | | | |
| Aggregated RCTs USc | 6–12 months | 271 | 0.65 ± 0.21 | [0.44 - 0.86] | 0.71 ± 0.20 | [0.51 - 0.91] |
| Aggregated RCTs VAS | 6–12 months | 271 | 54.26 ± 33.33 | [20.93–87.58] | 61.42 ± 20.07 | [41.35–81.49] |

*Note. M* = Mean; *SD* = Standard deviation; *DBT* = Dialectical Behaviour Therapy; RCT = Randomised Controlled trial; VAS - Visual Analogue Scale.

**Table 4**
Aggregated Benchmarks – DBT treatment efficacy benchmarks for adults (full-programme).

| Measure | *K* | *N* | *d*+ | $\sigma^2$ | *CV* | *Q* | *p* (*Q*) |
|---|---|---|---|---|---|---|---|
| *Aggregated treatment efficacy - benchmarks for RCT studies* | | | | | | | |
| BSL-23 | 2 | 240 | 0.70 | 0.006 | – | 0.02 | .901 |
| DERS | 2 | 114 | 1.25 | 0.013 | – | 1.31 | .253 |
| EQ-5D-USc | 4 | 271 | 0.28 | 0.004 | – | 0.92 | .819 |
| EQ-5D-VAS | 4 | 271 | 0.30 | 0.004 | – | 0.16 | .984 |
| *Aggregated treatment efficacy - benchmarks for effectiveness studies* | | | | | | | |
| BSL-23 | 4 | 117 | 0.882 | 0.011 | – | 33.10 | < 0.001 |
| DERS | 4 | 233 | 1.247 | 0.008 | – | 9.51 | <0.050 |
| *Treatment efficacy benchmarks for RCTs + Effectiveness* | | | | | | | |
| BSL-23 | 6 | 357 | 0.824 | 0.004 | 0.67 | 23.39 | <0.001 |
| DERS | 6 | 347 | 1.423 | 0.006 | 81.10 | 24.86 | <0.001 |
| EQ-5D-USc | 4 | 271 | 0.28 | 0.004 | – | 0.92 | .819 |
| EQ-5D-VAS | 4 | 271 | 0.30 | 0.004 | – | 0.16 | .984 |

*Note. DBT* = Dialectical Behaviour Therapy; *K* = number of samples included in analyses; *N* = sample size; *d* = unbiased pre–post effect size estimate; $\sigma^2$ = effect size variance; *CV* = Critical Value; *Q* = test of homogeneity; *p* = significance; BSL-23 = Borderline Symptom List; DERS = Difficulties in Emotion Regulation Scale; EQ-5D-USc: EQ-5D Utility Score; EQ-5D-VAS: EQ-5D Visual Analogue Scale.

studies.for adults receiving DBT full-programme and skills alone, in an outpatient setting. Thus, the following benchmarks for full-programme should be considered regarding BSL ($d_+$ = 0.824; CV = 0.67); DERS ($d_+$ = 1.423; CV = 81.10) and EQ-5D ($d_+$ = 1.423; $M_{post\text{-}treatment}$ = 0.71; $SD_{post\text{-}treatment}$= 0.2). To inform average means and standard deviations per type of trial at pre-treatment and post-treatment, consult Table 3.

**Table 5**
Aggregated benchmarks – DBT treatment efficacy benchmarks for adults (DBT skills).

| Measure | *K* | *N* | *d*+ | $\sigma^2$ | *CV* | *Q* | *Q* (*p*) |
|---|---|---|---|---|---|---|---|
| BSL-23 | 2 | 106 | 0.896 | 0.013 | 0.62 | 4.17 | .041 |
| DERS | 2 | 75 | 1.489 | 0.020 | 88.74 | 4.27 | <0.05 |

*Note. DBT* = Dialectical Behaviour Therapy; *K* = number of samples included in analyses; *N* = sample size; *d* = unbiased pre–post effect size estimate; $\sigma^2$ = effect size variance; *CV* = Critical Value; *Q* = test of homogeneity; *p* = significance; BSL-23 = Borderline Symptom List; DERS = Difficulties in Emotion Regulation Scale.

Additionally, the following benchmarks for DBT skills for adults in outpatient settings should be considered: BSL ($d_+$ = 0.896; CV = 0.62); DERS ($d_+$ = 1.489; CV = 88.74).

Fewer studies were retrieved with adolescents, allowing only for aggregated benchmarks for standard DBT (full-programme) for BSL ($d_+$ = 0.800; CV = 0.48).

It is important to take into consideration that, unlike the DERS and BSL, which are measures with a normal distribution, in the case of EQ-5D utility score, mean scores and effect sizes need to be considered with caution, and a usable critical value was not possible to retrieve using the calculation proposed by Minami et al. (2008).

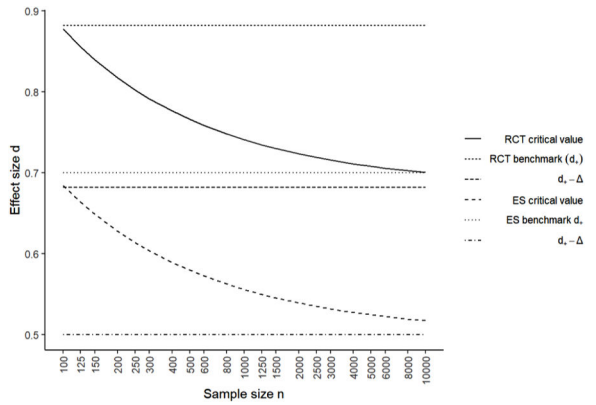


**Fig. 5.** BSL effect size critical values by study type and data sample size. *Note. RCT* = Randomised Controlled Trial; *ES* = Effect Size.

The benchmarks established in this article offer several advantages when reflecting on DBT delivery for, and treatment of, people with BPD. Firstly, benchmarks provide a standardised framework for assessing and evaluating progress throughout the therapy process and help set clear and measurable therapeutic aims. Secondly, establishing benchmarks enables regular assessment and monitoring of client progress. By comparing outcomes against predetermined targets, therapists can more objectively evaluate the effectiveness of their interventions, adjusting the treatment delivery and checking their adherence to the model if necessary, to deliver the best possible treatment.

Results from effectiveness studies potentially align more closely with the clients that clinicians may encounter in routine practice, whereas RCTs ideally provide benchmarks for when the highest quality of treatment is offered under tightly controlled conditions. Even so, clinicians should not be dissuaded from striving to attain results similar to those achieved in RCTs. Clinicians can consider these benchmarks as guidance, considering that the demographic and diagnostic mix (for example, baseline severity, socioeconomic status, and comorbidity) can vary widely across services.

To develop more representative and generalisable benchmarks in the future, establishing a further body of research using common instruments is a necessity, and we encourage researchers and DBT teams to use the instruments suggested in this article. In treatments of other mental health disorders, there has already been an effort to create a unified protocol to assess the effects of given treatments (Allen et al., 2008; Farchione et al., 2012). We believe this article makes a valuable contribution by supporting the systematic use of these instruments in the future. This, in turn, could facilitate a more consistent and standardised way of assessing DBT treatments, and their effectiveness.

### *Limitations and future directions*

In the future, we hope to be able to provide peer benchmarks in the context of an ongoing project to benchmark teams delivering DBT across

**Table 7**
Aggregated benchmarks – treatment efficacy benchmarks for adolescents (full-programme).

| Measure | *K* | *N* | $d_+$ | $\sigma^2$ | *CV* | *Q* | *p* (*Q*) |
|---|---|---|---|---|---|---|---|
| BSL-23 | 3 | 151 | 0.800 | 0.009 | 0.48 | 1.705 | .426 |

*Note. K* = number of studies; *N* = sample size; *d*+ = unbiased pre–post effect size estimate; $\sigma^2$ = effect size variance; *CV* = Critical Value; *Q* = test of homogeneity; *p* = significance; BSL-23 = Borderline Symptom List;.

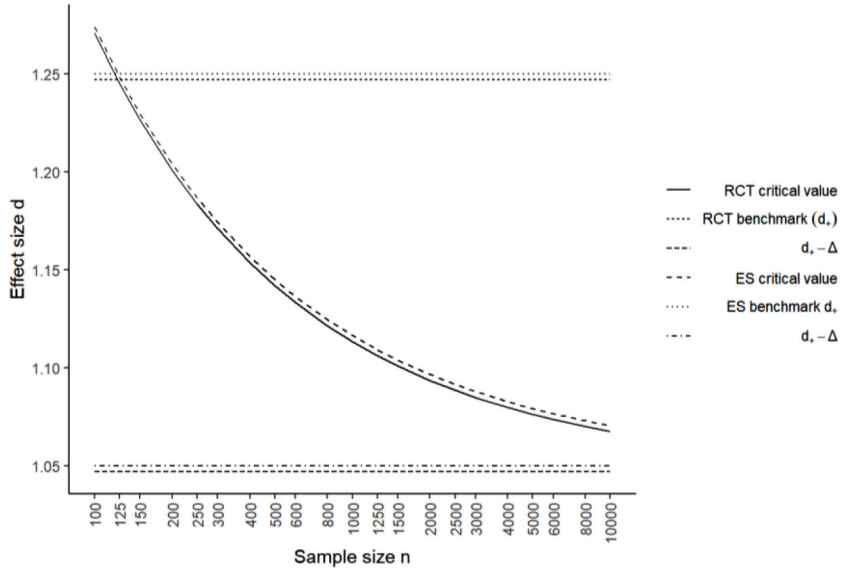


**Fig. 6.** DERS effect size critical values by study type and data sample size. *Note. RCT* = Randomised Controlled Trial; *ES* = Effect Size.

**Table 6**
Pre- and post-treatment outcomes of adolescent samples that received DBT full-programme treatment in the selected studies.

| | DBT details | *n* | Pre-treatment *M* ± *SD* | Pre-treatment *interval* | Post-treatment *M* ± *SD* | Post-treatment *interval* |
|---|---|---|---|---|---|---|
| *Borderline Symptoms List (BSL)* | | | | | | |
| Aggregated RCT+ Effectiveness | 16–24 weeks | 151 | 2.18 ± 0.87 | [1.31 – 3.05] | 1.47 ± 0.79 | [0.68 – 2.26] |
| *Difficulties in Emotion Regulation Scale (DERS)* | | | | | | |
| Berk et al., 2018 | 24 weeks | 24 | 3.33 ± 0.61 | [2.73 – 3.94] | 2.56± 0.80 | [1.75 – 3.36] |

*Note. M* = Mean, *SD* = Standard-deviation; DBT = Dialectical Behaviour Therapy; *n* = sample size;.

the UK and eventually be able to shed light on the similarities and differences between routine practice and the provided benchmarks.

Future RCTs and effectiveness studies should also seek to use an ITT approach and a standardised protocol to contribute to a common ground for researchers and clinicians. This should include the utilisation of consistent assessment tools on a global scale. Such an approach greatly simplifies the subsequent analysis and comparative studies and is a significant step towards benchmarking within the field of mental health.

Furthermore, looking at Devlin et al. (2020) chapter explaining different approaches to assess clinically significant changes using EQ-5D, we believe the use of an anchor measure for BPD, which could account for a relevant improvement in this population, could inform what would be a significant change. Future studies should investigate the use of BSL-23 as a possible anchor measure along with EQ-5D, to detect what could be consider a Minimum Important Difference (MID). Pickard et al. (2007) suggested using half of a standard deviation to calculate MID when using EQ-5D in a sample of cancer patients while selecting an anchor measure, and its methodology has been used in other clinical samples.

In terms of limitations, the studies we selected showed high heterogeneity, which stresses the need to be conservative in our conclusions. Moreover, we aggregated results from studies using different methodologies (ITT and completers) in order to establish more representative benchmarks, however, this comes at the cost of some level of accuracy that could have been attained with a more homogeneous sample. Participants who did not complete a given study may be systematically different from those who did, their exclusion can alter treatment effects, reducing the generalisability of the study's findings.

Benchmarks for adolescents were not possible to establish for the EQ-5D or DERS, because no studies were found using EQ-5D and only one study used DERS. The last can be used as a reference but not as a benchmark. In addition, we aggregated data from studies which investigated programmes of different lengths, essentially treating them as equal in terms of outcomes. Whilst in terms of clinical outcomes at the individual level this may be the case, it has a significant limitation. A programme that produces the same effect size in 6 months as a programme that is twice the length is twice as productive in health economic terms provided other things remain equal (team size and training). As DBT is a team-based treatment arguably treating clinical outcomes at the individual level may not be the best approach, although it is the approach that clinicians are most familiar with. Few of the studies that we reviewed and included in this paper systematically reported on 'team' as a variable or reported on any potential clustering of outcomes by site or in changes in outcome during the study due to learning effects in therapists, making it impossible for us to develop benchmarks that incorporated a 'team' factor. As we develop our peer-benchmarking platform we aim to address these important aspects of outcome in routine settings, in consultation with the teams using the platform, by systematically developing team 'productivity' measures, that incorporate treatment length, team resource and skill, to benchmark against.

## Conclusion

To assist teams delivering DBT in routine practice and to evaluate the impact of a national training programme on clinical outcomes, we searched the literature to compile benchmarks for three measures of outcome: the EQ-5D measuring health-related quality of life; the BSL, measuring borderline symptoms; and the DERS measuring difficulties in emotion regulation. We were able to compile aggregated benchmarks for teams working with adults delivering both comprehensive DBT (BSL $d_+ = 0.824$; CV = 0.67; DERS $d_+ = 1.423$; CV = 81.10; EQ-5D $d_+ = 1.423$; $M = 0.71 \pm 0.2$) and stand-alone skills training (BSL $d_+ = 0.896$; CV = 0.62; DERS $d_+ = 1.489$; CV = 88.74). Additionally, for teams delivering DBT full-programme to adolescents with BPD features, benchmarks were provided for BSL ($d_+ = 0.800$; CV = 0.48).

A smaller number of adolescent studies with higher heterogeneity limited the development of definitive benchmarks for all the instruments, although teams can still use the findings for comparison. Subsequent research will focus on using these benchmarks in routine practice to support the development of a peer network focused on the improvement of clinical outcomes and the development of peer benchmarks.

## Funding

This work was supported by the School of Psychology and Sport Science of Bangor University, British Isles DBT-Training (biDBT) and NHS England (formerly Health Education England (HEE).

## Declaration of competing interest

The author have no conflict of interest to disclose.

## Acknowledgments

We thank the School of Psychology and Sport Science, Bangor University, British Isles DBT-Training (biDBT) and NHS England (formerly Health Education England (HEE) for funding and supporting this project. We also thank the collaboration of Roland Sinaeve and Shelley McMain, who provided data from their studies, making a noteworthy contribution to the calculated benchmarks.

## References

Allen, L. B., McHugh, R. K., Barlow, D. H., & Barlow, D. H. (2008). Emotional disorders: A unified protocol. *Clinical handbook of psychological disorders: A step-by-step treatment manual* (4th ed., pp. 216–249). The Guilford Press.

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders (5th ed., text rev.)*.

Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health, 22*(4), 153–160. https://doi.org/10.1136/EBMENTAL-2019-300117

Barnicot, K., & Crawford, M. (2018). *Psychological Medicine Dialectical behaviour therapy v. mentalisation-based therapy for borderline personality disorder*. doi:10.1017/S0033291718002878.

Bayney, R. (2005). Benchmarking in mental health: An introduction for psychiatrists. *Advances in Psychiatric Treatment, 11*(4), 305–314. https://doi.org/10.1192/apt.11.4.305

Berk, M. S., Starace, N. K., Black, V. P., & Avina, C. (2018). Implementation of dialectical behavior therapy with suicidal and self-harming adolescents in a community clinic. *Archives of Suicide Research*. https://doi.org/10.1080/13811118.2018.1509750

Biskin, R. S. (2015). The lifetime course of borderline personality disorder. *Canadian Journal of Psychiatry, 60*(7), 303–308. https://doi.org/10.1177/070674371506000702

Bohus, M., Kleindienst, N., Limberger, M. F., Stieglitz, R. D., Domsalla, M., Chapman, A. L., et al. (2009). The short version of the Borderline Symptom List (BSL-23): Development and initial data on psychometric properties. *Psychopathology, 42*(1), 32–39. https://doi.org/10.1159/000173701

Buchholz, I., Janssen, M. F., Kohlmann, T., & Feng, Y. S. (2018). A systematic review of studies comparing the measurement properties of the three-level and five-level versions of the EQ-5D. *PharmacoEconomics, 36*(6), 645–661. https://doi.org/10.1007/s40273-018-0642-5

Carter, G.L., Willcox, C.H., Lewin, T.J., Conrad, A.M., & Bendit, N. (2010). Hunter DBT project: Randomized controlled trial of dialectical behaviour therapy in women with borderline personality disorder. 10.3109/00048670903393621, 44(2), 162–173. doi:10.3109/00048670903393621.

Chakhssi, F., Zoet, J. M., Oostendorp, J. M., Noordzij, M. L., & Sommers-Spijkerman, M. (2021). Effect of psychotherapy for borderline personality disorder on quality of life: A systematic review and meta-analysis. *Journal of Personality Disorders, 35*(2), 255–269. https://doi.org/10.1521/PEDI_2019_33_439

Chorpita, B. F., & Daleiden, E. L. (2014). Structuring the collaboration of science and service in pursuit of a shared vision. *Journal of Clinical Child and Adolescent Psychology: The Official Journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division, 53*(2), 323–338. https://doi.org/10.1080/15374416.2013.828297. *43*.

D'Aurizio, G., Di Stefano, R., Socci, V., Rossi, A., Barlattani, T., Pacitti, F., et al. (2023). The role of emotional instability in borderline personality disorder: A systematic review. *Annals of General Psychiatry, 22*(1). https://doi.org/10.1186/S12991-023-00439-0

Delgadillo, J., McMillan, D., Leach, C., Lucock, M., Gilbody, S., & Wood, N. (2014). Benchmarking routine psychological services: A discussion of challenges and

methods. *Behavioural and Cognitive Psychotherapy, 42*(1), 16–30. https://doi.org/10.1017/S135246581200080X
DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177–188. https://doi.org/10.1016/0197-2456(86)90046-2
Devlin, N., Parkin, D., & Janssen, B. (2020). Advanced topics. *Methods for analysing and reporting eq-5d data* (pp. 87–98). Springer International Publishing. https://doi.org/10.1007/978-3-030-47622-9_5
Eisen, S. V., & Dickey, B. (1996). Mental health outcome assessment: The new agenda. *Psychotherapy, 33*(2), 181–189. https://doi.org/10.1037/0033-3204.33.2.181
Eldh, A. C., Almost, J., Decorby-Watson, K., Gifford, W., Harvey, G., Hasson, H., et al. (2017). Clinical interventions, implementation interventions, and the potential greyness in between-a discussion paper. *BMC Health Services Research, 7*(17), 16. https://doi.org/10.1186/s12913-016-1958-5
EuroQol Research Foundation. (2018). *EQ-5D-3L user guide*. https://euroqol.org/publications/user-guides.
Farchione, T. J., Fairholme, C. P., Ellard, K. K., Boisseau, C. L., Thompson-Hollands, J., Carl, J. R., et al. (2012). Unified protocol for transdiagnostic treatment of emotional disorders: A randomized controlled trial. *Behavior Therapy, 43*(3), 666–678. https://doi.org/10.1016/J.BETH.2012.01.001
Gillespie, C., Joyce, M., Flynn, D., & Corcoran, P. (2019). Dialectical behaviour therapy for adolescents: A comparison of 16-week and 24-week programmes delivered in a public community setting. *Child and Adolescent Mental Health, 24*(3), 266–273. https://doi.org/10.1111/CAMH.12325
Gillespie, C., Murphy, M., & Joyce, M. (2022). Dialectical behavior therapy for individuals with borderline personality disorder: A systematic review of outcomes after one year of follow-up. *Journal of Personality Disorders, 36*(4), 431–454. https://doi.org/10.1521/pedi.2022.36.4.431
Glenn, C. R., & Klonsky, E. D. (2009). Emotion dysregulation as a core feature of borderline personality disorder. *Journal of Personality Disorders, 23*(1), 20–28. https://doi.org/10.1521/PEDI.2009.23.1.20
Gotham, H. J. (2006). Advancing the implementation of evidence-based practices into clinical practice: How do we get there from here?. In *Professional psychology: Research and practice, 37* pp. 606–613) American Psychological Association. https://doi.org/10.1037/0735-7028.37.6.606
Gratz, K. L., & Roemer, L. (2004). Multidimensional assessment of emotion regulation and dysregulation: Development, factor structure, and initial validation of the difficulties in emotion regulation scale. *Journal of Psychopathology and Behavioral Assessment, 26*(1), 41–54. https://doi.org/10.1023/B:JOBA.0000007455.08539.94
Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107. https://doi.org/10.2307/1164588
Heerebrand, S. L., Bray, J., Ulbrich, C., Roberts, R. M., Edwards, S., Services, D. N., et al. (2021). Effectiveness of dialectical behavior therapy skills training group for adults with borderline personality disorder. *Journal of Clinical Psychology*. https://doi.org/10.1002/jclp.23134
Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A, (Statistics in Society), 172*(1), 137–159. https://doi.org/10.1111/J.1467-985X.2008.00552.X
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
IsHak, W. W., Elbau, I., Ismail, A., Delaloye, S., Ha, K., Bolotaulo, N. I., et al. (2013). Quality of life in borderline personality disorder. *Harvard Review of Psychiatry, 21*(3), 138–150. https://doi.org/10.1097/HRP.0B013E3182937116
Joanna Briggs Institute. (2017). *The joanna briggs institute reviewers' manual 2017*. The Joanna Briggs Institute.
Kaufman, E. A., Xia, M., Fosco, G., Yaptangco, M., Skidmore, C. R., & Crowell, S. E. (2016). The Difficulties in Emotion Regulation Scale Short Form (DERS-SF): Validation and replication in adolescent and adult samples. *Journal of Psychopathology and Behavioral Assessment, 38*(3), 443–455. https://doi.org/10.1007/S10862-015-9529-3
Kells, M., Joyce, M., Flynn, D., Spillane, A., & Hayes, A. (2020). *Borderline personality disorder and emotion dysregulation*. 7(3). doi:10.1186/s40479-020-0119-y.
Kleindienst, N., Jungkunz, M., & Bohus, M. (2020). A proposed severity classification of borderline symptoms using the borderline symptom list (BSL-23). *Borderline Personality Disorder and Emotion Dysregulation, 7*(1), 1–11. https://doi.org/10.1186/S40479-020-00126-6/FIGURES/2
Kuehn, K. S., King, K. M., Linehan, M. M., & Harned, M. S. (2020). Modeling the suicidal behavior cycle: Understanding repeated suicide attempts among individuals with borderline personality disorder and a history of attempting suicide HHS Public Access. *Journal of Consulting and Clinical Psychology, 88*(6), 570–581. https://doi.org/10.1037/ccp0000496
Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2013). Why many clinical psychologists are resistant to evidence-based practice: Root causes and constructive remedies. *Clinical Psychology Review, 33*(7), 883–900. https://doi.org/10.1016/J.CPR.2012.09.008
Linehan, M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. Guilford Press.
Linehan, M. M. (2015). *DBT® skills training manual* (2nd ed.). Guilford Press.
Lloyd, R. C. (2004). *Quality health care: A guide to developing and using indicators* (1st ed.). Jones and Bartlett Publishers.
Lovaglio, P. G. (2012). Benchmarking strategies for measuring the quality of healthcare: Problems and prospects. *TheScientificWorldJournal, 2012*, Article 606154. https://doi.org/10.1100/2012/606154
Lyng, J., Swales, M. A., Hastings, R. P., Millar, T., & Duffy, D. J. (2020). Outcomes for 18 to 25-year-olds with borderline personality disorder in a dedicated young adult only DBT programme compared to a general adult DBT programme for all ages 18+. *Early Intervention in Psychiatry, 14*(1), 61–68. https://doi.org/10.1111/eip.12808
McMain, S. F., Chapman, A. L., Kuo, J. R., Dixon-Gordon, K. L., Guimond, T. H., Labrish, C., et al. (2022). The effectiveness of 6 versus 12 months of dialectical behavior therapy for borderline personality disorder: A noninferiority randomized clinical trial. *Psychotherapy and Psychosomatics, 91*(6), 382–397. https://doi.org/10.1159/000525102
McMain, S. F., Guimond, T., Barnhart, R., Habinski, L., & Streiner, D. L. (2017). A randomized trial of brief dialectical behaviour therapy skills training in suicidal patients suffering from borderline disorder. *Acta Psychiatrica Scandinavica, 135*(2), 138–148. https://doi.org/10.1111/ACPS.12664
McMain, S. F., Links, P. S., Gnam, W. H., Guimond, T., Cardish, R. J., Korman, L., et al. (2009). A randomized trial of dialectical behavior therapy versus general psychiatric management for borderline personality disorder. *The American Journal of Psychiatry, 166*(12), 1365–1374. https://doi.org/10.1176/APPI.AJP.2009.09010039
Mehlum, L., Tørmoen, A. J., Ramberg, M., Haga, E., Diep, L. M., Laberg, S., et al. (2014). Dialectical behavior therapy for adolescents with repeated suicidal and self-harming behavior: A randomized trial. *Journal of the American Academy of Child and Adolescent Psychiatry, 53*(10), 1082–1091. https://doi.org/10.1016/J.JAAC.2014.07.003
Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. S. (2008). Using clinical trials to benchmark effects produced in clinical practice. *Quality and Quantity, 42*(4), 513–525. https://doi.org/10.1007/S11135-006-9057-Z/METRICS
Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (2007). Benchmarks for psychotherapy efficacy in adult major depression. *Journal of Consulting and Clinical Psychology, 75*(2), 232–243. https://doi.org/10.1037/0022-006X.75.2.232
Mirkovic, B., Delvenne, V., Robin, M., Pham-Scottez, A., Corcos, M., & Speranza, M. (2021). Borderline personality disorder and adolescent suicide attempt: The mediating role of emotional dysregulation. *BMC psychiatry, 21*(1), 1–10. https://doi.org/10.1186/S12888-021-03377-X/FIGURES/2
Moroz, N., Moroz, I., & D'Angelo, M. S (2020). Mental health services in Canada: Barriers and cost-effective solutions to increase access. *Healthcare Management Forum, 33*(6), 282–287. https://doi.org/10.1177/0840470420933911
Neumann, A., van Lier, P. A. C., Gratz, K. L., & Koot, H. M. (2010). Multidimensional assessment of emotion regulation difficulties in adolescents using the difficulties in emotion regulation scale. *Assessment, 17*(1), 138–149. https://doi.org/10.1177/1073191109349579
Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews, 5*(1), 1–10. https://doi.org/10.1186/s13643-016-0384-4
Pickard, A. S., De Leon, M. C., Kohlmann, T., Cella, D., & Rosenbloom, S. (2007). Psychometric comparison of the standard EQ-5D to a 5 level version in cancer patients. *Medical Care, 45*(3), 259–263. https://doi.org/10.1097/01.MLR.0000254515.63841.81
R. Core Team. (2017). *A language and environment for statistical computing*.
Rizvi, S. L., Hughes, C. D., Hittman, A. D., & Oliveira, P. V. (2017). Can trainees effectively deliver dialectical behavior therapy for individuals with borderline personality disorder? Outcomes from a training clinic. *Journal of Clinical Psychology, 73*, 1599–1611. https://doi.org/10.1002/jclp.22467
Shadish, W. R., & Haddock, C. K. (1994). *The handbook of research synthesis*. Russell Sage Foundation.
Sinnaeve, R., Van Den Bosch, L. M. C., Hakkaart-Van Roijen, L., & Vansteelandt, K (2018). Effectiveness of step-down versus outpatient dialectical behaviour therapy for patients with severe levels of borderline personality disorder: A pragmatic randomized controlled trial. *Borderline Personality Disorder and Emotion Dysregulation, 5*(12). https://doi.org/10.1186/s40479-018-0089-5
Sloan, E., Hall, K., Moulding, R., Bryce, S., Mildred, H., & Staiger, P. K. (2017). Emotion regulation as a transdiagnostic treatment construct across anxiety, depression, substance, eating and borderline personality disorders: A systematic review. *Clinical Psychology Review, 57*, 141–163. https://doi.org/10.1016/J.CPR.2017.09.002
Stiglmayr, C., Stecher-Mohr, J., Meiβner, J., Spretz, D., Steffens, C., Roepke, S., et al. (2014). Effectiveness of dialectic behavioral therapy in routine outpatient care: The Berlin borderline study. *Borderline Personality Disorder and Emotion Dysregulation, 1*(20). https://doi.org/10.1186/2051-6673-1-20
Stoffers-Winterling, J. M., Storebø, O. J., Kongerslev, M. T., Faltinsen, E., Todorovac, A., Jørgensen, S., et al. (2022). Psychotherapies for borderline personality disorder: A focused systematic review and meta-analysis. *The British Journal of Psychiatry, 221*, 538–552. https://doi.org/10.1192/bjp.2021.204
Stoffers-Winterling, J. M., Völlm, B. A., Rücker, G., Timmer, A., Huband, N., & Lieb, K. (2012). Psychological therapies for people with borderline personality disorder. *The Cochrane Database of Systematic Reviews, 2012*(8). https://doi.org/10.1002/14651858.CD005652.PUB2. CD005652.
Swales, M. A., & Heard, H. L. (2017). *Dialectical behaviour therapy - The CBT distinctive features* (2nd ed.). Routledge.
Teachman, B. A., Drabick, D. A. G., Hershenberg, R., Vivian, D., Wolfe, B. E., & Goldfried, M. R. (2012). Bridging the gap between clinical research and clinical practice: Introduction to the special section. *Psychotherapy, 49*(2), 97–100. https://doi.org/10.1037/A0027346
Valentine, S. E., Bankoff, S. M., Poulin, R. M., Reidler, E. B., & Pantalone, D. W. (2015). The use of dialectical behavior therapy skills training as stand-alone treatment: A systematic review of the treatment outcome literature. *Journal of Clinical Psychology, 71*(1), 1–20. https://doi.org/10.1002/JCLP.22114

van Asselt, A. D. I., Dirksen, C. D., Arntz, A., Giesen-Bloo, J. H., & Severens, J. L. (2009). The EQ-5D: A useful quality of life measure in borderline personality disorder? *European Psychiatry : The Journal of the Association of European Psychiatrists, 24*(2), 79–85. https://doi.org/10.1016/J.EURPSY.2008.11.001

Videler, A. C., Hutsebaut, J., Schulkens, J. E. M., Sobczak, S., & van Alphen, S. P. J. (2019). A life span perspective on borderline personality disorder. *Current Psychiatry Reports, 21*(7), 51. https://doi.org/10.1007/s11920-019-1040-1

Walton, C. J., Bendit, N., Baker, A. L., Carter, G. L., & Lewin, T. J. (2020). A randomised trial of dialectical behaviour therapy and the conversational model for the treatment of borderline personality disorder with recent suicidal and/or non-suicidal self-injury: An effectiveness study in an Australian public mental health servic. *Australian and New Zealand Journal of Psychiatry, 54*(10), 1020–1034. https://doi.org/10.1177/0004867420931164

Washburn, M., Rubin, A., & Zhou, S. (2018). Benchmarks for outpatient dialectical behavioral therapy in adults with borderline personality disorder. *Research on Social Work Practice, 28*(8), 895–906. https://doi.org/10.1177/1049731516659363

Weersing, V. R., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology, 70*(2), 299–310. https://doi.org/10.1037/0022-006X.70.2.299

Weinberg, A., & Klonsky, E. D. (2009). Measurement of emotion dysregulation in adolescents. *Psychological Assessment, 21*(4), 616–621. https://doi.org/10.1037/A0016669

Zanarini, M. C., Horwood, J., Wolke, D., Waylen, A., Fitzmaurice, G., & Grant, B. F. (2011). Prevalence of DSM-IV borderline personality disorder in two community samples: 6330 english 11-year-olds and 34,653 American adults. *Journal of Personality Disorders, 25*(5), 607–619. https://doi.org/10.1521/pedi.2011.25.5.607