

Dialectical Behavior Therapy for Borderline Personality Disorder: A Meta-Analysis Using Mixed-Effects Modeling

Sören Kliem and Christoph Kröger
Technical University of Braunschweig

Joachim Kosfelder
University of Applied Sciences Düsseldorf

Objective: At present, the most frequently investigated psychosocial intervention for borderline personality disorder (BPD) is dialectical behavior therapy (DBT). We conducted a meta-analysis to examine the efficacy and long-term effectiveness of DBT. **Method:** Systematic bibliographic research was undertaken to find relevant literature from online databases (PubMed, PsycINFO, PsychSpider, Medline). We excluded studies in which patients with diagnoses other than BPD were treated, the treatment did not comprise all components specified in the DBT manual or in the suggestions for inpatient DBT programs, patients failed to be diagnosed according to the *Diagnostic and Statistical Manual of Mental Disorders*, and the intervention group comprised fewer than 10 patients. Using a mixed-effect hierarchical modeling approach, we calculated global effect sizes and effect sizes for suicidal and self-injurious behaviors. **Results:** Calculations of postintervention global effect sizes were based on 16 studies. Of these, 8 were randomized controlled trials (RCTs), and 8 were neither randomized nor controlled (nRCT). The dropout rate was 27.3% pre- to posttreatment. A moderate global effect and a moderate effect size for suicidal and self-injurious behaviors were found, when including a moderator for RCTs with borderline-specific treatments. There was no evidence for the influence of other moderators (e.g., quality of studies, setting, duration of intervention). A small impairment was shown from posttreatment to follow-up, including 5 RCTs only. **Conclusions:** Future research should compare DBT with other active borderline-specific treatments that have also demonstrated their efficacy using several long-term follow-up assessment points.

Keywords: meta-analyses, borderline personality disorder, dialectical behavior therapy, effectiveness study

Supplemental materials: <http://dx.doi.org/10.1037/a0021015.supp>

Dialectical behavior therapy (DBT; Linehan, 1993a, 1993b) is currently the most frequently investigated psychosocial intervention for borderline personality disorder (BPD). This comprehensive treatment program focuses on (a) promoting the motivation for change by detailed chain analyses, validation strategies, and management of reinforcement contingencies in individual therapy twice a week; (b) increasing target-oriented and appropriate behavior by teaching skills in a weekly group format training, fostering mindful attention and cognition, emotion regulation, acceptance of emotional distress, and interpersonal effectiveness; (c) ensuring the transfer of newly learned skills to everyday life by telephone coaching and case management; and (d) supporting therapists' motivation and skills with a weekly consultation team.

A treatment target hierarchy determines the problem focus of each session. Reduction of suicidal gestures and self-injurious

behaviors is given the highest priority in DBT (Linehan, 1993a, 1993b), considering that these behaviors predict completing suicide (Black, Blum, Pfohl, & Hale, 2004). Subsequently, patients were trained in skills geared to help them stay in outpatient therapy, followed by a reduction of comorbid Axis I disorders. Finally, quality-of-life issues or individual targets were addressed. Given that individuals with BPD are prone to frequent use of psychiatric facilities (Bender et al., 2001), the original outpatient model was modified for inpatient treatment (Swenson, Sanderson, Dulit, & Linehan, 2001). In previous studies (Bohus et al., 2004; Kleindienst et al., 2008; Kröger et al., 2006), findings support the assumption that a 3-month inpatient treatment program reduced self-rated general psychopathology, depression, anxiety, dissociation, and self-mutilating behavior at posttreatment and at follow-up.

The efficacy and effectiveness of DBT are summarized in several reviews (e.g., Lieb, Zanarini, Schmahl, Linehan, & Bohus, 2004; Oldham, 2006). The current *Cochrane Review* relies on the results from some (Koons et al., 2001; Linehan, Armstrong, Suarez, Allmon, & Heard, 1991; Linehan et al., 2002, 1999; van den Bosch, Koeter, Stijnen, Verheul, & van den Brink, 2005) but not all available randomized controlled trials (RCTs; Clarkin, Levy, Lenzenweger, & Kernberg, 2007; Linehan, Comtois, Murray, et al., 2006; Linehan, McDavid, Brown, Sayrs, & Gallop, 2008; McMMain et al., 2009; Simpson et al., 2004). It also includes a trial comprising psychodynamic techniques in individual therapy and a six-session group therapy focusing on significant others (Turner,

Sören Kliem and Christoph Kröger, Department of Psychology, Technical University of Braunschweig, Braunschweig, Germany; Joachim Kosfelder, Department of Social Sciences and Cultural Studies, University of Applied Sciences Düsseldorf, Düsseldorf, Germany.

The first authors contributed equally to this work.

Correspondence concerning this article should be addressed to Christoph Kröger, Department of Psychology, Technical University of Braunschweig, Humboldtstraße 33, 38106 Braunschweig, Germany. E-mail: c.kroeger@tu-bs.de

2000). In three studies ($n = 155$) in the *Cochrane Review*, there was no difference in the dropout rate compared with treatment as usual (TAU). No further integrated effect sizes were reported. The authors concluded that individuals with BPD “may be amenable to talking/behavioural treatments” (Binks et al., 2006, p. 20). In a current meta-analysis, based on 13 exclusively randomized controlled studies, a mean effect of 0.58 (95% CI [0.38, 0.77]) was reported for the effectiveness of DBT (Öst, 2008). However, treatment studies focusing on disorders other than BPD were also included, and two recent RCTs evaluating the effectiveness of DBT for BPD (Clarkin et al., 2007; McMains et al., 2009) were not considered for this meta-analysis. Furthermore, no follow-up data were reported. Therefore, there is a need for conducting a further meta-analysis focusing on clinical trials using DBT in the treatment of BPD.

The largest possible body of primary studies should be used in meta-analysis, to prevent limited generalizability and selection bias (Sica, 2006). There is an ongoing debate as to which design provides the best evidence that a treatment works (e.g., Barlow, 1996; Seligman, 1995; Westen, Novotny, & Thompson-Brenner, 2004). If only RCTs are included in meta-analyses, high internal validity can be expected due to the design controlling for factors that have an impact on outcome outside the treatment in question. Effectiveness studies conducted under the condition of clinical routine offer more external validity. However, those studies are mostly noncontrolled trials (nRCTs).

The aim of this meta-analytic review is to examine (a) the effectiveness pre- to posttreatment in general and on suicidal and self-injurious behaviors and, for the first time, (b) the long-term effectiveness of DBT for BPD. Trying to include all available data, also from nRCTs, we used hierarchical linear modeling (HLM) to account for the nested data structure encountered in meta-analyses. Compared with conventional analysis of data, this procedure relies on Bayesian estimation of the overall effect size, which has been shown to be more appropriate in meta-analyses with a small number of studies (DuMouchel, 1994).

Method

Literature Research and Selection of Studies

Systematic bibliographic research was undertaken to find relevant literature from online databases (PubMed, PsycINFO, PsychSpider, Medline) using the following keywords: *dialectical behavior therapy*, *DBT*, and their German equivalences. Additional articles were found through references in reviews and empirical studies, as well as by Internet search and contact with research groups. Studies published up until the end of October 2009 were surveyed. Two independent raters (S. K. and C. K.; the latter is supervisor for behavior therapy and a DBT therapist, board-certified by the German Association for Dialectical Behavior Therapy) extracted the articles ($\kappa = .93, p < .001$). Those studies were excluded in which (a) individuals other than BPD patients were treated; (b) patients failed to be diagnosed according to the *Diagnostic and Statistical Manual of Mental Disorders*; (c) the treatment was conducted not using four components (individual therapy, group format training, consultation team, telephone or staff coaching) with contents specified in the manual (Linehan, 1993a, 1993b) or in the suggestions for inpatient DBT programs (Swen-

son et al., 2001);¹ and (d) the intervention group comprised fewer than 10 patients, as a sample size of 10 or more is recommended for adequate precision in calculating the effect size variance (Hedges & Olkin, 1985).

Although we used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses standards (Moher, Liberati, Tetzlaff, Altman, & the PRISMA Group, 2009), a randomized controlled design was not required for any study to be included in the present meta-analysis. In including both types of study (RCT plus nRCT), two strategies have been proposed. First, we estimate effect sizes by including all RCTs and adding nRCTs afterward. This procedure immediately reveals bias tendencies. Second, sensitivity analyses, in contrast, are proposed for the entire analysis (RCTs plus nRCTs), to analyze the differential effect size estimation of RCTs and nRCTs. We used a likelihood ratio test to compare the results from the two models. The first model excludes the moderator effect; the second model estimates all effects. For each model, a deviance statistic is computed, and the difference between the deviance statistics is used to compare the model fits. A significant likelihood ratio test means that there is a difference between the RCTs and the nRCTs.

Calculating Effect Sizes

For RCTs, between-groups effect sizes were calculated according to Hedges and Olkin (1985) by dividing the difference between group means at postintervention by the pooled standard deviation between groups (Hedges's g ; all formulae are presented in the Appendix). Odds ratios were used to calculate effect sizes for categorical data (Fleiss, 1994; see Appendix). The log-odds ratio was transformed into Hedges's g (Haddock, Rindskopf, & Shadish, 1998; Hasselblad & Hedges, 1995; see Appendix). Other effect measures, such as the product-moment correlation, were transformed into Hedges's g (Rosenthal, 1994; see Appendix).

For nRCTs, within-group effect sizes were calculated by standardizing pre- and posttreatment and pre-follow-up mean differences for each intervention group by the standard deviation of the difference (Hartmann & Herzog, 1995; Johnson, 1989; see Appendix). To obtain these standard deviations, we estimated the correlations from repeated measures t statistics or single-group repeated measures analyses of variance for the relevant time points (DeCoster, 2009; Morris & DeShon, 2002; Rosenthal, 1994; see Appendix).

Effect sizes were corrected for possible bias due to small sample sizes (Hedges & Olkin, 1985; Hunter & Schmidt, 1994; Matt & Cook, 1994; Shadish & Haddock, 1994; see Appendix). According to Hedges (1981), neglecting this adjustment would cause an overestimation of the integrated effects. The effect sizes post-follow-up are considered as effect gain (Becker, 1988) by subtracting the post-effect sizes from the follow-up effect sizes (i.e., respective subtraction of the pre- and posttreatment effect sizes from the pre-follow-up effect sizes). The effect size variance for

¹ To ensure generalizability of the findings, variability of weekly treatment dose and duration of the intervention could differ from the manual (Linehan, 1993a). In further analyses, we included a moderator to control the efficacy of DBT as a function of the intervention's duration. See supplemental materials for a table summarizing the components of included studies.

the between-groups effect sizes has been calculated according to Hedges and Olkin (1985) by adding up the reciprocal value of the harmonic mean of both sample sizes with the squared effect size divided by the doubled sum of both sample sizes (Hedges, Cooper, & Bushman, 1998; see Appendix). When calculating the within-group effect size by standardizing at the standard deviation of the difference, the effect size variance can be derived directly from the between-groups effect size variance (Gibbons, Hedecker, & Davis, 1993; see Appendix). For the interpretation of the estimation of the effect size, Cohen (1988) suggested a classification whereby values of effect sizes were rated as small (>0.2), medium (>0.5), and large (>0.8).

Statistical Model

The mixed-effects hierarchical model assumes that the intervention effects are to be drawn from a whole population of effect sizes (Konstantopoulos, 2006). To integrate the effect sizes based on the small study number, we applied HLM, allowing for appropriate analysis of the meta-analytic data. HLM uses information from all the studies to obtain an empirical Bayes estimate of each study's effects (Raudenbush & Bryk, 2002). Each study effect size was shrunk toward the grand mean considering the sample size (Kreft & de Leeuw, 1998; see Appendix). Shrinkage is large when sample sizes are small, but it is small when large sample sizes are used. The Bayesian calculations have proven more stable in meta-analyses with a small number of studies (DuMouchel, 1994).

In the present study, the calculation was modeled on three levels (see Appendix). Level 1, the bottom level, represents individuals nested within outcome measures. The variance of the individuals was estimated with the variance of the effect sizes. Level 2 captures outcome measures nested within studies. Level 3, the uppermost level, represents the differences between studies. If there is only one outcome measure, the three-level model is reduced to a two-level model. The analysis starts with an unconditional model without including any explanatory variable at either level. This enables an estimation of the overall effect size and the amount of heterogeneity within levels. Homogeneity is tested with the statistic H , which provides an estimate of the extent to which sample effects deviate from the grand mean, weighted by inverse of the variance (Hedges & Olkin, 1985; Raudenbush & Bryk, 2002; see Appendix). With the low number of primary studies and the associated high beta error taken into account, a conditional model was generally applied, even though the test for heterogeneity was nonsignificant. A conditional model includes predictor variables that might account for observed variance. SPSS Version 16.0 and HLM 6 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004) were used for the meta-analytic integration.

Moderators

To control for potential confounding factors, we included several moderator variables. First, we rated the methodological quality of the primary studies using the checklist of Downs and Black (1998), allowing an assessment of RCTs as well as nRCTs. The scale assesses the methodological quality with four subscales: reporting, external validity, internal validity, and power. The maximum number of points is 32 for RCTs and 28 for nRCTs. This instrument has high internal consistency ($KR20 = 0.89$), high

test-retest reliability ($r = .88$), and good interrater reliability ($r = .75$). The literature has frequently shown how studies with low methodological quality tend to yield extreme results (Egger, Jüni, Bartlett, Hohenstein, & Sterne, 2003; MacLehose et al., 2000; Moher et al., 1998). To assess the interrater reliability of the scale, a postdoctoral student in clinical psychology received 3 hr of training in the use of the scale by one of the authors (C. K.).² The postdoctoral student was blind (blackened text) to the investigators as well as to whether the study was excluded. She rated the included studies and 25% of the excluded studies. Then the ratings were compared with those of the first author (S. K.). Second, the original DBT outpatient concept was adapted to an inpatient setting (Swenson et al., 2001). This conceptual change will be considered by entering a dichotomous moderator. Third, the moderator duration of the intervention is supposed to illustrate a linear or quadratic trend of the efficacy of DBT as a function of the duration of the intervention. Similarly, the period between postintervention and follow-up is supposed to illustrate the decline of achieved improvement. We used the last follow-up assessment point. Fourth, the percentage of dropouts is a potential moderator variable, which enabled us to assess whether trials with high attrition rates (i.e., narrowed effective populations of high responders) differ in outcome from trials keeping more participants at postintervention (Rüsch et al., 2008).

Multiple Outcome Measures

There are two general approaches to dealing with multiple outcome measures within studies. In the single-value approach, each study is represented by a single value. For example, this can be the average measurement per study (Durlak, 2000). This procedure does not perform very well with respect to recovering the true effect size (Bijmold & Pieters, 2001). In the complete-set approach, all outcomes are individually included in the analysis, using all available information. The most commonly used procedure is to incorporate the values of all measurements within studies and to treat these as independent replications (Bijmold & Pieters, 2001). In this case, studies with many outcomes may have a larger effect on the results of meta-analysis than studies with few measurements (Rosenthal, 1991). Hence, we count the outcomes within each study as an independent weighted replication within an HLM framework, using all available information (Raudenbush & Bryk, 2002). This procedure turns out to be sensible in a Monte Carlo comparison (Bijmold & Pieters, 2001). To determine one main outcome measure for every study did not seem sensible to us, because (a) it implies a significant loss of information and a loss in construct validity (Bijmold & Pieters, 2001), (b) even the developer of DBT refers to several main outcome measures in her primary studies, and (c) the heterogeneity of the reported outcome measures (34 measures) does not allow for reasonable determination of one specific main outcome measure.

Specific Outcomes

Suicidal and self-injurious behaviors. The following measures were applied to assess suicidal and self-injurious behavior:

² We would like to thank Melanie Vonau, Technical University of Braunschweig.

Lifetime Parasuicide Count (Linehan & Comtois, 1996), Overt Aggression Scale–Modified (Coccaro, Harvey, Kupsaw-Lawrence, Herbert, & Bernstein, 1991), and Suicide Attempt Self-Injury Interview (formerly called the Parasuicide History Interview; Linehan, Comtois, Brown, Heard, & Wagner, 2006). Furthermore, the rates of patients engaging in self-injurious behaviors and suicide attempts were included. Given that self-injurious behaviors are not normally distributed, we applied odds ratios to calculate effect sizes (Fleiss, 1994; see Appendix). The log-odds ratios were transformed into Hedges's g (Haddock et al., 1998; Hasselblad & Hedges, 1995; see Appendix). Additionally, chi-square values were converted into Hedges's g (Rosenthal, 1994; see Appendix).

Dropout. Because the maintaining of therapy takes the second priority on the target hierarchy, we calculated effect sizes based on odds ratios (Fleiss, 1994) between the dropout rates of DBT and control conditions from pretreatment to post-treatment.

Results

The initial search yielded 75 studies that reported empirical evidence on DBT. Figure 1 shows a summary of the study selection process. After filtering according to the inclusion criteria, 26 trials were included. Table 1 describes all 26 studies that were included in the meta-analysis. When results based on the same sample were reported in several studies, sample data were included only once, resulting in a total sample of 16.

Of all the studies that were selected for the analysis, eight were classified as controlled by randomization (Clarkin et al., 2007; Koons et al., 2001; Linehan et al., 1991, 2002, 1999; Linehan, Comtois, Murray, et al., 2006; McMMain et al., 2009; van den Bosch et al., 2005), seven were included as neither randomized nor controlled (Comtois, Elwood, Holdcraft, Smith, & Simpson, 2007; Friedrich, Gunia, & Huppertz, 2003; Höschel, 2006; Kröger et al., 2006; Linehan et al., 2008; Prendergast & McCausland, 2007; Simpson et al., 2004), and one was not randomized but controlled (Bohus et al., 2004). Because this study was not controlled at follow-up (Kleindienst et al., 2008), it was added to the nRCT group of studies.

Specific characteristics for some included studies need to be addressed. Noteworthy are the first studies by Linehan and colleagues, which included data from two cohorts from pretreatment to posttreatment (Linehan et al., 1991) as well as to follow-up (Linehan, Heard, & Armstrong, 1993) and only secondary outcome measures from the second cohort from pretreatment to posttreatment (Linehan, Tutek, Heard, & Armstrong, 1994). For further analysis, only those outcome measures were used that were based on both cohorts. Furthermore, DBT was compared in two RCTs with borderline-specific treatments (Clarkin et al., 2007; McMMain et al., 2009). Using one of the other borderline-specific treatments as a control condition might lead to a considerable underestimation of effect sizes for DBT compared with nonspecific interventions (e.g., TAU). Hence, we computed a sensitivity analysis using a dichotomous moderator characterizing both these RCTs. Finally, two studies included a pure DBT control group and an intervention group with DBT plus pharmacological treatment (Line-

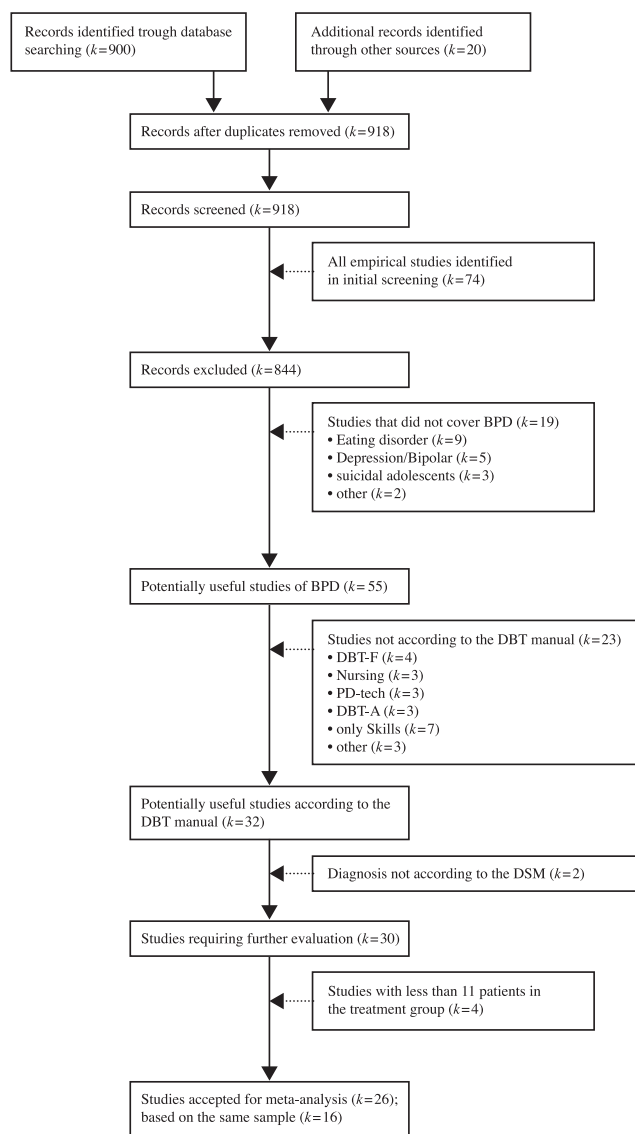


Figure 1. Selection process for studies to be included in the meta-analyses. BPD = borderline personality disorder; DBT = dialectical behavior therapy; DBT-F = dialectical behavior therapy for forensic patients; PD-tech = psychodynamic techniques; DBT-A = dialectical behavior therapy for suicidal adolescents. DSM = *Diagnostic and Statistical Manual of Mental Disorders*.

han et al., 2008; Simpson et al., 2004) classified as nRCTs. The medication dosage could not be controlled in this meta-analysis due to a lack of information in most of the primary studies. Therefore, effect sizes of both groups have been pooled. However, adding a drug to investigate a potentially enhancing effect of this drug might have a higher impact than just allowing the patients to continue to use some of the drugs they are already consuming. Therefore, we included a dichotomous moderator characterizing both studies with add-on pharmacological treatments to control for possible bias and computed a sensitivity analysis by using a likelihood ratio test.

Table 1
Studies Included in a Meta-Analysis of Dialectical Behavior Therapy (DBT) for Borderline Personality Disorders (BPDs)

Study	Design	Assessment points	Sample size	Mean age (SD), gender	Exclusion criteria	Inclusion criteria	Dropout	Method	Setting
Bohus et al., 2004, 2000; Kleinendienst et al., 2008	nRCT	0, 4 (post), 12, 24 (FU) month	DBT = 31	29.6 (7.5), 100% female	Axis I: schizophrenia, bipolar I disorder, substance abuse; Axis II: mental retardation	BPD, female, one suicide attempt or a minimum of two non-suicidal self-injuries within the last 2 years	Post = 25.8%, FU = 29%	ITT-LOCF	In
Clarkin et al., 2004, 2007	RCT	0, 4, 8, 12 (post) month	DBT = 30, ST = 30, TFP = 30	30.9 (7.85), 96% female	Axis I: psychotic disorder, bipolar I disorder; substance dependence, delirium, dementia, amnesic disorder, other cognitive disorders; Axis II: mental retardation	BPD	Post = 31.1%, post-DBT = 43.3%	ITT-LOCF	Out
Comtois et al., 2007	nRCT	0, 12 (post) month	DBT = 38	34 (range: 19–54), 96% female	Axis I: substance dependence; Axis II: mental retardation	BPD, extensive suicide attempts, crisis service history	Post = 37%	Completers	Out
Fassbinder et al., 2007; Kröger et al., 2006	nRCT	0, 3 (post), 15, 30 (FU) month	DBT = 50	30.5 (7.7), 88% female	Axis I: schizophrenia, bipolar I disorder, substance abuse, substance dependence, dementia; Axis II: mental retardation; Axis III: current symptoms	BPD, age > 18	Post = 12%, FU = 40%	ITT-LOCF	In
Friedrich et al., 2003	nRCT	0, 3, 6, 9, 12 (post) month	DBT = 33	33.4 (8.8), 91% female	Axis I: acute psychosis	BPD	Post = 8%	Completers	Out
Harned et al., 2008; Linehan, Comtois, Murray, et al., 2006	RCT	0, 12 (post), 24 (FU) month	DBT = 52, CTBE = 49	29.3 (7.5), 100% female	Axis I: schizophrenia, schizoaffective disorder, psychotic disorder not otherwise specified, bipolar disorder; Axis II: mental retardation; Axis III: seizure disorder; Other: requiring medication, a mandate to treatment	BPD, female, recent and recurrent self-injury	Post = 11.8%, FU = 19.8%, post-DBT = 3.8%, FU-DBT = 11.5%	ITT-RRM	Out
Hörschel, 2006	nRCT	0, 2, 8, 12 (post) weeks	DBT = 24	28.3 (6.86), 87.5% female	Axis I: schizophrenic disorder, substance dependence; Axis II: mental retardation, antisocial personality disorder	BPD	Post = 4.2%	Completers	In
Koons et al., 2001	RCT	0, 3, 6 (post) month	DBT = 14, TAU = 14	34.5 (7.5), 100% female	Axis I: schizophrenic disorder, bipolar disorder, substance dependence; Axis II: mental retardation, antisocial personality disorder	BPD, female veterans	Post = 28.8%, post-DBT = 28.8%	Completers	Out
Linehan et al., 1991, 1993, 1994	RCT	0, 4, 8, 12 (post), 18, 24 (FU) month	DBT = 22, TAU = 22	26.7 (7.8), 100% female	Axis I: Schizophrenic disorder, bipolar disorder, substance dependence; Axis II: mental retardation	BPD, female 18 < age < 45, suicide attempt in the past 8 weeks, one other in the past 5 years	Post = 6.8%, FU = 18.2%, post-DBT = 9.1%, FU-DBT = 18.2%	ITT-LOCF	Out

Table 1 (continued)

Study	Design	Assessment points	Sample size	Mean age (SD), gender	Exclusion criteria	Inclusion criteria	Dropout	Method	Setting
Linehan et al., 2002	RCT	0, 4, 8, 12 (post), 16 (FU) month	DBT = 11, CVT + 12 = 12	36.1 (7.3), 100% female	Axis I: psychotic disorder, bipolar disorder; Axis II: mental retardation; Axis III: seizure disorder	BPD, female, current opiate dependence	Post = 20.8%, FU = 20.8%, post-DBT = 36.4%, FU-DBT = 36.4%	ITT-LOCF	Out
Linehan et al., 2008	nRCT	0, 7, 12, 21 (post) weeks	DBT = 24	26.8 (9.0), 100% female	Axis I: psychotic disorder, bipolar I disorder, major depressive disorder with psychotic features, substance dependence; Axis II: mental retardation; Axis III: seizure disorder, pregnant, breastfeeding, planning to become pregnant; Other: episode of self-inflicted injury in the 8 weeks prior to the screening interview	BPD, female, OAS-M ≤ 6	Post = 33%	ITT-RRM	Out
Linehan et al., 1999	RCT	0, 4, 8, 12 (post), 16 (FU) month	DBT = 12, TAU = 16	30.4 (6.6), 100% female	Axis I: psychotic disorder, bipolar disorder; Axis II: mental retardation	BPD female, current drug dependence	Post = 61.1%, FU = 64.3%, post-DBT = 33.3%, FU-DBT = 41.7%	ITT-LOCF	Out
McMain et al., 2009	RCT	0, 4, 8, 12 (post) month	DBT = 90, GPM = 90	29.4 (9.1), 90% female	Axis I: psychotic disorder, bipolar I disorder, delirium, dementia; Axis II: of substance dependence in the preceding 30 days, having a medical condition that precluded psychiatric medications	BPD, 18 < age < 60, at least two episodes of suicidal or non-suicidal self-injurious episodes in the past 5 years, at least one of which was in the 3 months preceding enrollment	Post = 38.3%, post-DBT = 38.9%	ITT-LOCF	Out
Prendergast & McCausland, 2007	nRCT	0, 6 (post) month	DBT = 16	36.35 (7.42), 100% female	Axis I: psychotic disorder, substance dependence	BPD, female	Post = 31%	SC	Out
Simpson et al., 2004	nRCT	0, 12 (post) weeks	DBT = 25	35.3 (10.1), 100% female	Axis I: schizophrenia, bipolar I disorder, substance dependence; Axis III: seizure disorder, pregnant, lactating; Other: unwilling to use effective birth control, unstable medical conditions, monoamine oxidase inhibitor treatment in the last 2 weeks, a previous adequate trial of fluoxetine	BPD	Post = 20%	SC	In
van den Bosch et al., 2005, 2002, 2001; Verheul et al., 2003	RCT	Baseline, 0, 52 (post), 78 (FU) weeks	DBT = 27, TAU = 31	34.9 (7.7), 100% female	Axis I: psychotic disorder, bipolar disorder; Axis II: mental retardation	BPD, female	Post = 17.2%, FU = 24.1%, post-DBT = 14.6%, FU-DBT = 25.9%	ITT-RRM	Out

Note. nRCT = nonrandomized and noncontrolled trial; RCT = randomized controlled trial; post = postintervention; FU = follow-up; ITT = intention to treat; LOCF = last observation carried forward; In = inpatient setting; Out = outpatient setting; ? = no information available; ST = supportive treatment; TFP = transference-focused psychotherapy; SC = statistical control; TAU = therapy as usual; CTBE = community therapy by experts; RRM = random-effects regression model; CVT + 12 = comprehensive validation plus 12-step therapy; OAS-M = Overt Aggression Scale-Modified; GPM = general psychiatric management.

Global Effect From Preintervention to Postintervention

Calculations of postintervention global effect sizes were based on all studies. The total number of treated patients was 794; of these, 217 (27.3%) dropped out between pretreatment and post-treatment. In the DBT condition, there were 499 patients, of whom 123 (24.7%) dropped out between preintervention and postintervention.

Effect of RCTs. Analyzing only RCTs ($k = 8$; number of treated patients = 553, after dropout = 391; number of patients treated with DBT = 258, after dropout = 190) resulted in an effect size estimation of 0.39, 95% CI [0.10, 0.68], $t(7) = 2.59$, $p = .036$ (two-tailed).

Including the moderator that considers the impact of borderline-specific controlled RCTs (Clarkin et al., 2007; McMain et al., 2009) yields an effect size estimation of 0.51, 95% CI [0.38, 0.64], $t(6) = 8.05$, $p > .001$ (two-tailed). The moderator effect, interpreted as the difference between the effect of the borderline-specific controlled RCTs ($k = 2$) and RCTs without that control condition ($k = 6$), was -0.50 , 95% CI $[-0.63, -0.37]$, $t(6) = -7.68$, $p > .001$ (two-tailed). A likelihood ratio test indicated a significant improvement of the model quality when comparing the unconditional and conditional model, $\chi^2(1) = 12.06$, $p < .001$.

Combined effect of RCTs and nRCTs. There was no evidence for bias tendencies for nRCTs, $\chi^2(1) = 0.006$, $p = .938$. Because sensitivity analysis indicates a bias tendency for add-on pharmacological treatments, $\chi^2(1) = 7.20$, $p = .007$, those studies ($k = 2$) were excluded. Adding nRCTs ($k = 6$; number of treated patients = 154, after dropout = 126) resulted in an effect size estimation of 0.44, 95% CI [0.27, 0.61], $t(13) = 5.18$, $p < .001$ (two-tailed). There was significant unexplained variance on Level 3 ($H = 109.81$, $df = 13$, $p < .001$). Figure 2A shows the ordinary least squares estimates of the global effects in a forest plot.

Including the moderator that considers the impact of borderline-specific controlled RCTs yields an effect size estimation of 0.50, 95% CI [0.43, 0.57], $t(12) = 14.7$, $p > .001$ (two-tailed). The moderator effect was -0.49 , 95% CI $[-0.56, 0.42]$, $t(12) = -13.5$, $p > .001$ (two-tailed). A likelihood ratio test indicated a significant improvement of the model quality, $\chi^2(1) = 20.87$, $p < .001$. There was no significant unexplained variance on Level 3 ($H = 17.89$, $df = 12$, $p = .119$).

Model fit. To test the homogeneity of the error variance, we plotted residual versus predicted values on Level 1. The assumption of a normal distribution of the unexplained variance was checked with a normal Q-Q plot and a Kolmogorov-Smirnov test ($z = 1.05$, $N = 118$, $p = .208$, two-tailed). No violation of the model assumptions was found.

Tests for publication bias. To reduce the file-drawer effect, we tried to identify unpublished studies. Figure 2B shows in a funnel plot the relationship between study effects and sample sizes. The assumption of a normal distribution of samples' effect sizes was checked with a normal Q-Q plot and a Kolmogorov-Smirnov test ($z = 0.71$, $N = 14$, $p = .627$, two-tailed; Begg & Berlin, 1988; Greenhouse & Iyengar, 1994). Hence, there was no evidence for publication bias in the conventional inspection. As another test for publication bias, we assessed the fail-safe number according to Rosenthal (1979). Eight nonpublished RCTs with effect sizes of 0 and a sample size of 35 in the experimental group and in the

control group had to be included in the analysis in order to reduce the obtained effect size of 0.39 of the eight RCTs to a small effect size of about 0.2. Seventeen additional nRCTs with an effect size of 0 and a sample size of 29 in the experimental group had to be conducted in order to decrease the effect size of 0.44 of the 14 included studies to 0.2.

Moderators

The mean methodological quality of the studies, based on the checklist developed by Downs and Black (1998), was 22.0 points ($SD = 2.0$). According to MacLehose (2000), nRCTs with low methodological quality might tend to overestimate the effect of the intervention. Therefore, the methodological quality of RCTs and nRCTs was compared. A Mann-Whitney U test showed no significant difference between the means ($z = -0.13$, $p = .461$, one-tailed). We determined the interrater reliability of the two blind raters by using the intraclass correlation coefficient for rater consistency in a two-way mixed model, with raters as fixed and studies as random factors (Rustenbach, 2003). The subscales of the checklist by Downs and Black (1998) were included as dependent variables. The intraclass correlation coefficient for single measures was .97 with a 95% CI [0.95, 0.98]. The mean dropout was 25.8% ($SD = 15.6\%$; minimum = 4.2%, maximum = 61.1%). No significant differences emerged between the mean dropout rates of RCTs and nRCTs (Mann-Whitney U test, $z = -0.32$, $p = .805$, two-tailed). The mean duration of the intervention was 40.0 weeks ($SD = 17.2$; minimum = 12 weeks, maximum = 52 weeks). The mean number of summarized effects was 8.6 ($SD = 4.1$; minimum = 1, maximum = 16). No impact of moderators emerged in our analyses with the exception of the moderator considering the impact of RCTs with borderline-specific controlled treatments.³ A likelihood ratio test comparing the model that included only this moderator with the model that included all moderators was nonsignificant, $\chi^2(5) = 2.47$, $p = .781$. No difference was found between a linear ($p = .486$) and quadratic ($p = .507$) function for the relationship between time and global study effect sizes.

Effect of Treatment as Usual

Analyzing the pre- to posttreatment effect sizes of TAU of five studies (Bohus et al., 2004; Koons et al., 2001; Linehan et al., 1991, 1999; van den Bosch et al., 2005; number of treated patients with TAU = 83, after dropout = 59) resulted in an effect size estimation of 0.11, 95% CI $[-0.20, 0.42]$, $t(4) = 0.67$, $p = .541$ (two-tailed). There was significant unexplained variance on Level 2 ($H = 11.27$, $df = 4$, $p = .023$). No violation of the model assumptions was found (Kolmogorov-Smirnov test, $z = 0.63$, $p = .999$, two-tailed).

Specific Outcome: Suicidal and Self-Injurious Behaviors

Fifteen studies report effects for self-injurious behaviors (except for Kröger et al., 2006). Four studies (Höschel, 2006; Linehan et

³ See supplemental materials for results.

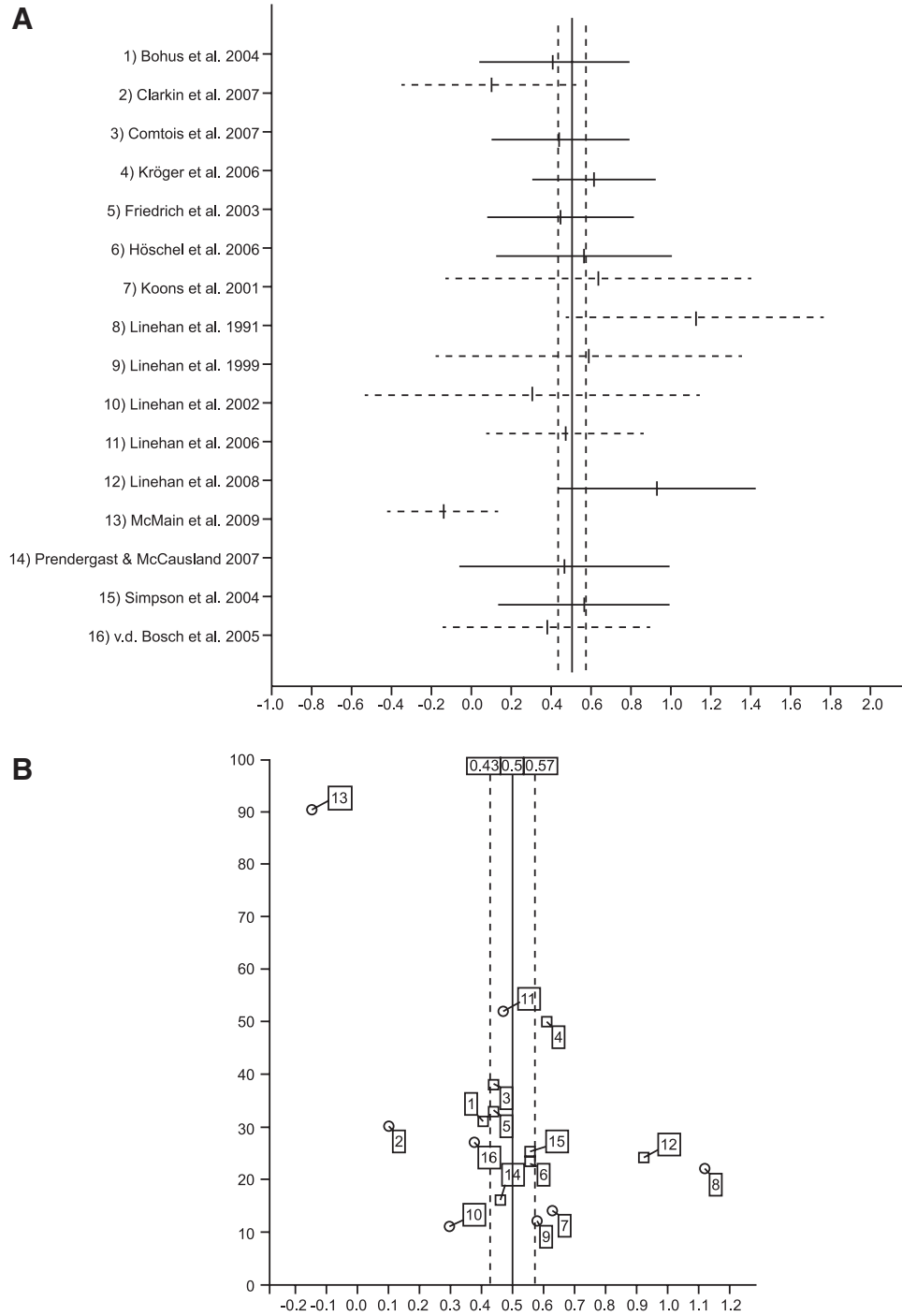


Figure 2. (A) The forest plot of ordinary least squares estimate for the moderated global effect sizes (95% confidence interval). Dashed lines denote randomized controlled trials (RCTs). (B) The funnel plot shows the relationship between study effects (x-axis) and sample sizes of the dialectical behavior therapy treatment group (y-axis). Circles denote RCTs.

al., 2002, 1999; Simpson et al., 2004) did not report rates to calculate odds ratios; hence, these studies were excluded from analyses. Because suicidal and self-injurious behaviors were not inclusion criteria for the remaining studies, we used a dichotomous moderator characterizing studies that examine samples with high

rates of self-injurious behaviors ($k = 6$). The number of treated patients was 643; of these, 183 (28.5%) dropped out between pretreatment and posttreatment. There were 377 patients treated by DBT, of which 103 (27.3%) dropped out between preintervention and postintervention.

Effect of RCTs. Analyzing only RCTs ($k = 6$) resulted in an effect size estimation of 0.23, 95% CI [-0.00, 0.46], $t(5) = 1.93$, $p = .110$ (two-tailed).

Including the moderator that considers the impact of borderline-specific controlled RCTs (Clarkin et al., 2007; McMain et al., 2009) yields an effect size estimation of 0.60, 95% CI [0.49, 0.71], $t(4) = 10.61$, $p > .001$ (two-tailed). The moderator effect was -0.56 , 95% CI [-0.67, -0.45], $t(4) = -9.91$, $p > .001$ (two-tailed). A likelihood ratio test for improvement of the model quality from an unconditional to a conditional model was significant, $\chi^2(1) = 10.68$, $p < .001$.

Combined effect of RCTs and nRCTs. Adding nRCTs ($k = 5$; number of treated patients = 142, after dropout = 99) resulted in an effect size estimation of 0.37, 95% CI [0.17, 0.57], $t(10) = 3.59$, $p = .006$ (two-tailed). There was no evidence for bias tendencies for nRCTs, $\chi^2(1) < 0.28$, $p = .597$. There was significant unexplained variance on Level 3 ($H = 45.55$, $df = 10$, $p < .001$). Figure 3A shows the ordinary least squares estimates of the study effects in a forest plot.

Including the moderator that considers the impact of borderline-specific controlled RCTs (Clarkin et al., 2007; McMain et al., 2009) yields an effect size estimation of 0.56, 95% CI [0.52, 0.60],

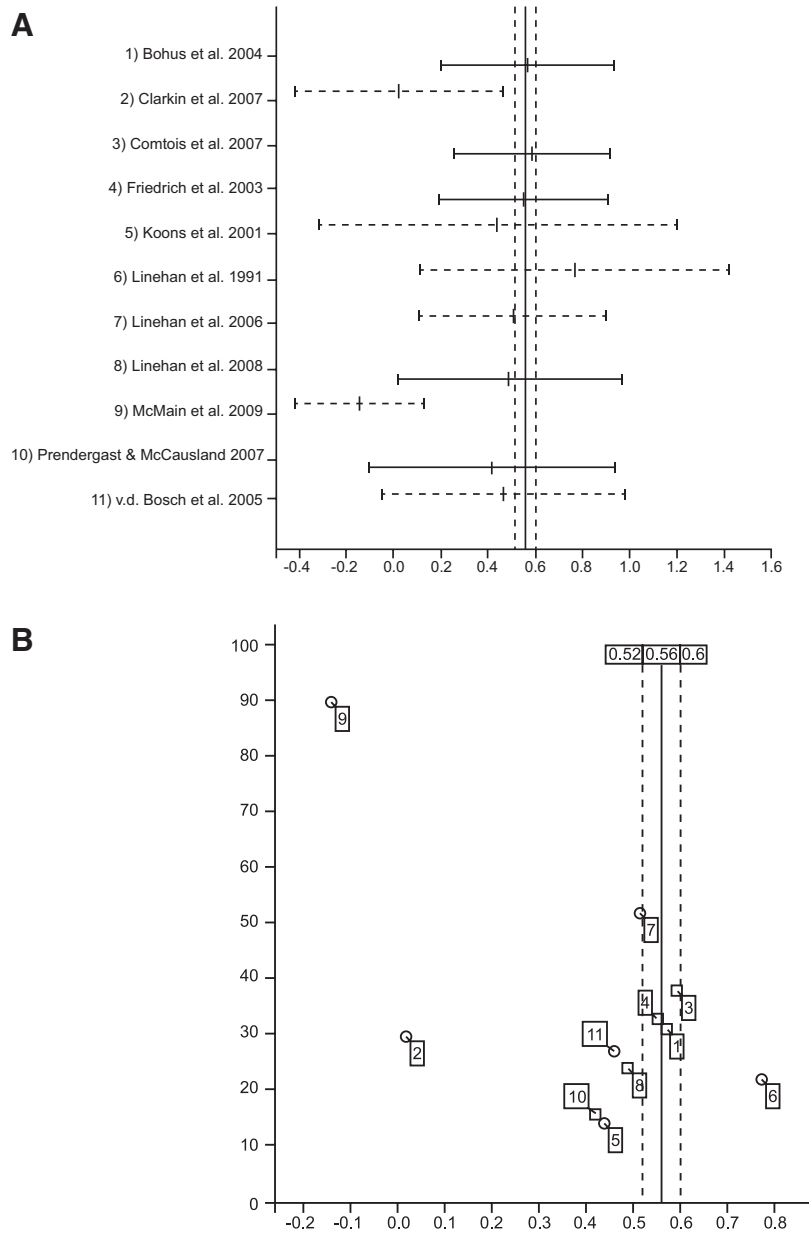


Figure 3. (A) The forest plot of ordinary least squares estimate for the moderated effect sizes (95% confidence interval) of suicidal and self-injurious behaviors. Dashed lines denote randomized controlled trials (RCTs). (B) The funnel plot shows the relationship between study effects (x-axis) and sample sizes of the dialectical behavior therapy treatment group (y-axis). Circles denote RCTs.

$t(9) = 27.04, p > .001$ (two-tailed). The moderator effect was $-0.52, 95\% \text{ CI} [-0.56, -0.48], t(9) = -23.47, p > .001$ (two-tailed). A likelihood ratio test for improvement of the model quality from an unconditional to a conditional model was significant, $\chi^2(1) = 19.5, p < .001$. There was no significant unexplained variance on Level 3 ($H = 3.3, df = 9, p = .951$).

Model fit. Again, no violation of the model assumptions was found (Kolmogorov–Smirnov test, $z = 0.66, N = 22, p = .729$, two-tailed). A likelihood ratio test for improvement of the model quality from an unconditional to a conditional model was nonsignificant, $\chi^2(5) = 0.82, p = .976$. No impact of moderators emerged in our analyses,⁴ nor was any difference found between a linear ($p = .230$) and quadratic ($p = .221$) function for the relationship between time and global study effect sizes.

Tests for publication bias. Figure 3B shows in a funnel plot the relationship between study effects and sample sizes. Again, there was no evidence for publication bias (Kolmogorov–Smirnov test, $z = 1.09, N = 11, p = .186$, two-tailed). One nonpublished RCT with an effect size of 0 and a sample size of 42 in the experimental group and in the control group had to be included in the analysis, reducing the obtained effect size of 0.23 of the six RCTs to a small effect size of 0.2. Nine additional nRCTs with an effect size of 0 and sample size of 34 in the experimental group had to be conducted in order to decrease the effect size of 0.37 of the 11 included studies to an effect size of 0.2.

Dropout Rate

Comparison of dropout rates between DBT and control conditions ($k = 8$) resulted in a global effect size of 0.03, $95\% \text{ CI} [-0.46, 0.52], t(7) = 0.11, p = .910$ (two-tailed). There was significant unexplained variance on Level 2 ($H = 45.6, df = 7, p < .001$). No violation of the model assumption was found (Kolmogorov–Smirnov test, $z = 0.81, p = .450$, two-tailed).

Global Effect Post-Follow-Up

Calculation of post-follow-up global effect sizes was based on seven studies, of which five were categorized as RCTs (Linehan, Comtois, Murray, et al., 2006; Linehan et al., 2002, 1993, 1999; van den Bosch et al., 2005), and two were included as neither randomized nor controlled (Fassbinder et al., 2007; Kleindienst et al., 2008). The total number of patients was 336; of these, 94 (28.0%) dropped out between pretreatment and follow-up. There were 205 patients treated with DBT, of which 55 (26.8%) dropped out between preintervention and follow-up.

Effect of RCTs. Analyzing only RCTs ($k = 5$; number of treated patients = 255, after dropout to postintervention = 203, after dropout to follow-up = 190; number of patients treated with DBT = 124, after dropout to postintervention = 108, after dropout to follow-up = 98) resulted in an effect size estimation of $-0.20, 95\% \text{ CI} [-0.25, -0.15], t(4) = -8.37, p < .001$ (two-tailed), without evidence for unexplained variance on Level 3 ($H = 1.12, df = 4, p = .891$).

Combined effect of RCTs and nRCTs. Adding nRCTs ($k = 2$; number of treated patients = 81, after dropout to postintervention = 67, after dropout to follow-up = 52) resulted in an effect size estimation of $-0.05, 95\% \text{ CI} [-0.22, 0.12], t(6) = -0.59, p = .578$ (two-tailed). A significant likelihood ratio test, $\chi^2(1) =$

$9.33, p = .003$, indicates the bias tendencies for nRCT. Taking this positive result of the sensitivity analysis into account, we analyzed a model fit including only RCTs.

Model fit. To test the homogeneity of the error variance, we plotted residual versus predicted values on Level 1. The assumption of normal distribution was checked for the error variance (Kolmogorov–Smirnov test, $z = 0.67, p = .964$, two-tailed) and samples' effect size (Kolmogorov–Smirnov test, $z = 0.39, p = .991$, two-tailed). Figure 4A shows the ordinary least squares estimates of the posttreatment to follow-up effect in a forest plot. Figure 4B shows in a funnel plot the relationship between study effects and sample sizes.

Moderators

The mean interval from posttreatment to follow-up was 32.4 weeks ($SD = 18.4$; minimum = 16 weeks, maximum = 52 weeks). The mean methodological quality was 23.8 points ($SD = 1.9$; minimum = 21, maximum = 26). The mean dropout from preintervention to follow-up was 29.45% ($SD = 19.6\%$; minimum = 18%, maximum = 64%) and from postintervention to follow-up was 7.7% ($SD = 4.6\%$; minimum = 0%, maximum = 12%). The mean number of summarized effects was 3.8 ($SD = 3.4$; minimum = 1, maximum = 8). No impact of moderators was found. Again, no difference was found between a linear ($p = .431$) and quadratic function ($p = .555$) for the relationship between time and posttreatment to follow-up effect estimation. A likelihood ratio test for improvement of the model quality from an unconditional to a conditional model was nonsignificant, $\chi^2(2) = 0.21, p = .900$.

Discussion

The present meta-analysis found a moderate effect size for DBT in the treatment of BPD patients. However, this holds true when we compare DBT with TAU, comprehensive validation plus 12-step therapy and community therapy by experts, whereas effect sizes decrease to small when DBT is compared with borderline-specific treatments. Although we found no evidence for the relative efficacy of DBT compared with other borderline-specific treatments, it is important to note that finding no significant differences between treatments in a study does not allow a conclusion that the treatments are equivalent. For example, transference-focused psychotherapy was a treatment condition in one of the RCTs with borderline-specific treatment (Clarkin et al., 2007). In this study, transference-focused psychotherapy was consistently related only to the reduction in aggression compared with DBT. Hence, there might be different impacts of treatment approaches on the heterogenic symptoms of BPD individuals. However, no follow-up data were reported for both RCTs with borderline-specific controlled treatments yet.

Adding nRCTs to the calculation of effect sizes yields smaller moderated confidence intervals, indicating a better estimation of the true effect size. This result was obtained when summarizing all reported outcome measures as well as when focusing on the reduction of suicidal and self-injurious behaviors in particular.

⁴ See supplemental materials for results.

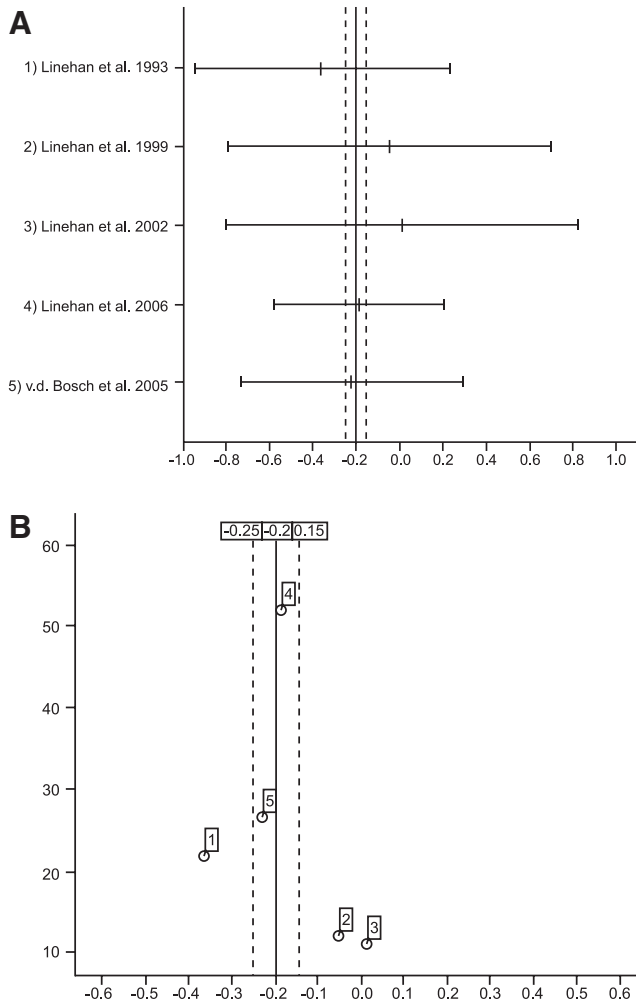


Figure 4. (A) The forest plot of ordinary least squares estimate for the posttreatment to follow-up effect sizes (95% confidence interval). Dashed lines denote randomized controlled trials (RCTs). (B) The funnel plot shows the relationship between study effects (x-axis) and sample sizes of the dialectical behavior therapy treatment group (y-axis). Circles denote RCTs.

These findings support the assumption that DBT is effective in clinical practice as well. Our results also confirm similar overall effect sizes recently reported by Öst (2008), who did not select studies for diagnosis and used a fixed-effects model. However, contrary to one of the main aims of DBT (Linehan, 1993a), no significant difference in the dropout rates between DBT and control conditions was found.

The moderated global effect of DBT decreased at follow-up, indicating that more research is needed to improve the transfer in daily life. Given that only five RCTs were included, however, this finding should be interpreted with caution. Both excluded nRCTs (Fassbinder et al., 2007; Kleindienst et al., 2008) were conducted in an inpatient setting in Germany, collecting data at long-term follow-up assessment points (30 and 20 months, respectively). In both nRCTs, patients received treatment (community therapy by experts or TAU) during the follow-up period, which might have a positive impact on global effect sizes at follow-up.

Generalizability

All selected studies obtained at least satisfactory scores in the quality ratings; therefore, the overall methodological quality can be considered adequate. Confidence intervals of the effects were narrow as a result of the Bayesian estimation. A possible bias caused by nonpublished studies (publication bias) is supported neither by nonexplained variance nor by the inspection of funnel plots. Because only eight RCTs and eight nRCTs could be integrated, the most stable calculation was used. Sensitivity analyses found no bias in effect size estimation when nRCTs were integrated as well. However, it has to be noted that the sample sizes in the DBT treatment groups of RCTs were mainly small ($n < 30$), with the exception of Linehan et al. (2006; $n = 52$), Clarkin et al. (2007; $n = 30$), and McMain et al. (2009; $n = 90$).

To ensure treatment integrity, we included only studies reporting four components of DBT (individual therapy, group format training, consultation team, telephone or staff coaching). Because several adaptations were reported (e.g., outpatient and inpatient setting, duration of group sessions, frequency of telephone and staff coaching), we assume that our findings indicate that DBT is a robust treatment approach across clinical practice. All RCTs were conducted with the supervision or cooperation of Marsha Linehan as the developer of DBT—with one exception (Clarkin et al., 2007), in which acknowledged experts supervised the three treatment conditions. Noteworthy are the large effect sizes in the developer's study (Linehan et al., 1991; see Figure 2B, Study 8; Figure 3B, Study 6), indicating a possible bias due to an overlap of therapist and researcher team (Linehan et al., 1994). Adherence scales were used in five studies (Clarkin et al., 2007; Koons et al., 2001; Linehan, Comtois, Murray, et al., 2006; McMain et al., 2009; van den Bosch et al., 2005), whereas none of the nRCTs applied an adherence scale. With the exception of one study (Prendergast & McCausland, 2007), all nRCTs were conducted with the supervision or cooperation of known DBT experts.

Limitations

There are limitations that should be considered in interpreting the results. We have been able to include only published reports, which might result in a loss of studies that have not been published (e.g., doctoral dissertations or oral presentations). Moreover, trials using only components of DBT (e.g., skills training) or component-control designs could not be integrated.

Although we tried to control for moderating effects, the number of moderators being analyzed is limited. Further aspects of the included studies might be considered as having an impact on effect sizes. For example, we assume high comorbidity rates for BPD with Axis I disorders, even though relevant data were missing in a number of studies (e.g., Bohus et al., 2004; Friedrich et al., 2003; Linehan et al., 1991; van den Bosch et al., 2005). Mood, anxiety, and eating disorders have the highest comorbidity rates in the outpatient (Zimmerman & Mattia, 1999) and inpatient settings (Zanarini et al., 1998). These comorbid disorders might complicate the planning and implementation of treatment and might exert a negative impact on effect sizes.

Because of the heterogeneity of BPD symptoms, a variety of measures that were mainly nonspecific to BPD were used. Overall, 34 different measures were reported, 18 of which were self-report

measures. On average, about eight effect sizes per study were summarized. Yet the risk remains that specific effects cannot be detected because of chosen measures. For instance, there is evidence that self-report instruments tend to obtain more valid information on experiential symptoms at the criteria level, whereas interviews tend to obtain more valid information on behavioral symptoms (Hopwood et al., 2008). It was already shown in the first study (Linehan et al., 1991) that self-rated emotional impairment (depression, hopelessness, suicide ideation, or questioning the reason for living) did not improve. In the area of suicidal and self-injurious behaviors, a major focus of DBT, the more effective interviews assess the type and frequency of those behaviors (e.g., Suicide Attempt Self-Injury Interview; Linehan, Comtois, Brown, et al., 2006). Therefore, we assume that the moderate global effect size might not be attributed primarily to methodological difficulties, as both experiential and behavioral symptoms were obtained by two types of measures (i.e., self-reports vs. interviews that provide valid information).

In addition, we were unable to control the impact of utilization of psychiatric services or other psychosocial treatments in parallel to DBT (e.g., Linehan et al., 1991; Linehan, Comtois, Murray, et al., 2006; McMain et al., 2009) as well as treatments during the follow-up period (e.g., Fassbinder et al., 2007; Kleindienst et al., 2008; Linehan et al., 1993).

Implications

A moderate effect size also points to the potential for improving the treatment. Although several studies have evaluated the DBT's effectiveness, there has been less emphasis on the processes and mechanism of change. Even though there is no evidence based on the data reviewed in this meta-analysis, relying on process research (e.g., Shearin & Linehan, 1992) or adding and evaluating treatment modules targeting specific syndromes (e.g., the consequences of sexual and physical abuse; Harned & Linehan, 2008) might improve standard DBT and make it even more effective.

To facilitate outcome comparisons in different domains for future studies, we suggest that the scientific community come to an agreement about a core battery of assessment measures. Some suggestions were already made for personality disorders in general (Strupp, Horowitz, & Lambert, 1997), including assessment of different perspectives and dimensions of change. There are appropriate measures for general BPD symptoms (e.g., Borderline Personality Disorder Severity Index–Version IV; Arntz et al., 2003), and for suicidal and self-injurious behaviors in particular (e.g., Suicide Attempt Self-Injury Interview; Linehan, Comtois, Brown, et al., 2006), which could be used as a standard in future research.

Future studies should compare active borderline-specific treatments that have also demonstrated their efficacy: for example, schema-focused therapy (Giesen-Bloo et al., 2006), transference-focused therapy (Clarkin et al., 2007), psychoanalytic therapy (Bateman & Fonagy, 1999, 2008), and general psychiatric management (Kolla et al., 2009; McMain et al., 2009). Considering that different effects for treatment conditions were revealed in such comparison studies, several long-term (>3 years) follow-up assessment points need to be conducted in order to rule out differential strength of competing treatment approaches.

References

References marked with an asterisk indicate studies included in the meta-analysis.

- Arntz, A., van den Hoorn, M., Cornelis, J., Verheul, R., van den Bosch, W. M. C., & de Bie, A. J. H. T. (2003). Reliability and validity of the Borderline Personality Disorder Severity Index. *Journal of Personality Disorders, 17*, 45–59.
- Barlow, D. H. (1996). Health care policy, psychotherapy research, and the future of psychotherapy. *American Psychologist, 51*, 1050–1058.
- Bateman, A., & Fonagy, P. (1999). Effectiveness of partial hospitalization in the treatment of borderline personality disorder: A randomized controlled trial. *American Journal of Psychiatry, 156*, 1563–1569.
- Bateman, A., & Fonagy, P. (2008). 8-year follow-up of patients treated for borderline personality disorder: Mentalization-based treatment versus treatment as usual. *American Journal of Psychiatry, 165*, 631–638.
- Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257–278.
- Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society: Series A. Statistics in Society, 151*, 419–463.
- Bender, D. S., Dolan, R. T., Skodol, A. E., Sanislow, C. A., Dyck, I. R., McGlashan, T. H., . . . Gunderson, J. G. (2001). Treatment utilization by patients with personality disorders. *American Journal of Psychiatry, 158*, 295–302.
- Bijmold, T. H. A., & Pieters, R. G. M. (2001). Meta-analysis in marketing when studies contain multiple measurements. *Marketing Letters, 12*, 157–169.
- Binks, C. A., Fenton, M., McCarthy, L., Lee, T., Adams, C. E., & Duggan, C. (2006). Psychological therapies for people with borderline personality disorder. *Cochrane Database Systematic Reviews* (1, Article CD005652). doi:10.1002/14651858.CD005652
- Black, D. W., Blum, N., Pfohl, B., & Hale, N. (2004). Suicidal behavior in borderline personality disorder: Prevalence, risk factors, prediction, and prevention. *Journal of Personality Disorders, 18*, 226–239.
- *Bohus, M., Haaf, B., Simms, T., Limberger, M. F., Schmahl, C., Unkel, C., Lieb, K., & Linehan, M. M. (2004). Effectiveness of inpatient dialectical behavioral therapy for borderline personality disorder: A controlled trial. *Behaviour Research and Therapy, 42*, 487–499.
- *Bohus, M., Haaf, B., Stiglmayr, C. H., Pohl, U., Böhme, R., & Linehan, M. M. (2000). Evaluation of inpatient dialectical-behavioral therapy for borderline personality disorder—A prospective study. *Behaviour Research and Therapy, 38*, 875–887.
- *Clarkin, J. F., Levy, K. N., Lenzenweger, M. F., & Kernberg, O. F. (2004). The Personality Disorders Institute/Borderline Personality Disorder Research Foundation randomized control trial for borderline personality disorder: Rationale, methods, and patient characteristics. *Journal of Personality Disorders, 18*, 52–72.
- *Clarkin, J. F., Levy, K. N., Lenzenweger, M. F., & Kernberg, O. F. (2007). Evaluating three treatments for borderline personality disorder: A multiwave study. *American Journal of Psychiatry, 164*, 922–928.
- Coccaro, E. F., Harvey, P. D., Kupsaw-Lawrence, E., Herbert, J. L., & Bernstein, D. P. (1991). Development of neuropharmacologically based behavioral assessment of impulsive aggressive behavior. *Journal of Neuropsychiatry and Clinical Neurosciences, 3*, S44–S51.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- *Comtois, K. A., Elwood, L., Holdcraft, L. C., Smith, W. R., & Simpson, T. L. (2007). Effectiveness of dialectical behavior therapy in a community mental health center. *Cognitive and Behavioral Practice, 14*, 406–414.
- DeCoster, J. (2009). Meta-analysis notes. Retrieved from <http://www.stat-help.com/notes.html>

- Downs, S., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health, 52*, 377–384.
- DuMouchel, W. (1994). *Hierarchical Bayes linear models for meta-analysis* (Tech. Rep. No. 27). Research Triangle Park, NC: National Institute of Statistical Sciences.
- Durlak, J. A. (2000). How to evaluate a meta-analysis. In D. Drotar (Ed.), *Handbook of research in pediatric and clinical child psychology: Practical strategies and methods* (pp. 395–407). New York, NY: Kluwer Academic/Plenum.
- Egger, M., Jüni, P., Bartlett, C., Hohenstein, F., & Sterne, J. (2003). How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. *Health Technology Assessment, 7*, 1–76.
- *Fassbinder, E., Rudolf, S., Bussiek, A., Kröger, C., Arnold, R., Gregersen, W., . . . Schweiger, U. (2007). Effektivität der dialektischen Verhaltenstherapie bei Patienten mit Borderline-Persönlichkeitsstörung im Langzeitverlauf. Eine 30-Monats-Katamnese nach stationärer Behandlung [Effectiveness of dialectical behavior therapy for patients with borderline personality disorder in the long-term course: A 30-month-follow-up after inpatient treatment]. *Psychotherapie, Psychosomatik, Medizinische Psychologie, 57*, 1–9.
- Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York, NY: Russell Sage Foundation.
- *Friedrich, J., Gunia, H., & Huppertz, M. (2003). Evaluation eines ambulanten Netzwerks für Dialektisch Behaviorale Therapie [Evaluation of an outpatient dialectical behavioral therapy network]. *Verhaltenstherapie und Verhaltensmedizin, 24*, 289–306.
- Gibbons, R. D., Hedecker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics, 18*, 271–279.
- Giesen-Bloo, J. G., van Dyck, R., Spinhoven, P., van Tilburg, W., Dirksen, C., van Asselt, T., . . . Arntz, A. (2006). Outpatient psychotherapy for borderline personality disorder: Randomized trial of schema-focused therapy vs. transference-focused psychotherapy. *Archives of General Psychiatry, 63*, 649–658.
- Greenhouse, J. B., & Iyengar, S. (1994). Sensitivity analysis and diagnostics. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 383–398). New York, NY: Russell Sage Foundation.
- Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods, 3*, 339–353.
- *Harned, M. S., Chapman, A. L., Dexter-Mazza, E. T., Murray, A., Comtois, K. A., & Linehan, M. M. (2008). Treating co-occurring Axis I disorders in recurrently suicidal women with borderline personality disorder: A 2-year randomized trial of dialectical behavior therapy versus community treatment by experts. *Journal of Consulting and Clinical Psychology, 76*, 1068–1075.
- Harned, M. S., & Linehan, M. M. (2008). Integrating dialectical behavior therapy and prolonged exposure to treat co-occurring borderline personality disorder and PTSD: Two case studies. *Cognitive and Behavioral Practice, 15*, 263–276.
- Hartmann, A., & Herzog, T. (1995). Varianten der Effektstärkenberechnung in Meta-Analysen: Kommt es zu variablen Ergebnissen? [Calculating effect size by varying formulas: Are there varying results?]. *Zeitschrift für Klinische Psychologie und Psychotherapie, 24*, 337–343.
- Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117*, 167–178.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128.
- Hedges, L. V., Cooper, H., & Bushman, B. J. (1998). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin, 111*, 188–194.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
- Hopwood, C. J., Morey, L. C., Edelen, M. O., Shea, M. T., Grilo, C. M., Sanislow, C. A., . . . Skodol, A. E. (2008). Comparison of interview and self-report methods for the assessment of borderline personality disorder criteria. *Psychological Assessment, 20*, 81–85.
- *Höschel, K. (2006). Dialektisch Behaviorale Therapie der Borderline Persönlichkeitsstörung in der Regelversorgung—Das Saarbrücker DBT-Modell [Dialectical behavioral therapy for borderline personality disorder in standard medical care: The Saarbrücken Treatment Program]. *Verhaltenstherapie, 16*, 17–24.
- Hunter, J. E., & Schmidt, F. L. (1994). Correcting for sources of artificial variation across studies. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 323–336). New York, NY: Russell Sage Foundation.
- Johnson, B. T. (1989). *Dstat: Software for the meta-analytic review of research literatures*. Hillsdale, NJ: Erlbaum.
- *Kleindienst, N., Limberger, M. F., Schmahl, C., Steil, R., Ebner-Priemer, U. W., & Bohus, M. (2008). Do improvements after inpatient dialectical behavioral therapy persist in the long term? A naturalistic follow-up in patients with borderline personality disorder. *Journal of Nervous and Mental Disease, 196*, 847–851.
- Kolla, N. J., Links, P. S., McMains, S., Streiner, D. L., Cardish, R., & Cook, M. (2009). Demonstrating adherence to guidelines for the treatment of patients with borderline personality disorder. *Canadian Journal of Psychiatry, 54*, 181–189.
- Konstantopoulos, S. (2006). *Fixed and mixed effects models in meta-analysis* (Discussion Paper No. 2198). Bonn, Germany: Institute for the Study of Labor.
- *Koons, C. R., Robins, C. J., Tweed, J. L., Lynch, T. R., Gonzales, A. M., Morse, J. Q., . . . Bastian, L. A. (2001). Efficacy of dialectical behavior therapy in women veterans with borderline personality disorder. *Behavior Therapy, 32*, 371–390.
- Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.
- *Kröger, C., Schweiger, U., Sipos, V., Arnold, R., Kahl, K. G., Schunert, T., . . . Reinecker, H. (2006). Effectiveness of dialectical behaviour therapy for borderline personality disorder in an inpatient setting. *Behaviour Research and Therapy, 44*, 1211–1217.
- Lieb, K., Zanarini, M. C., Schmahl, C., Linehan, M. M., & Bohus, M. (2004). Borderline personality disorder. *Lancet, 364*, 453–461.
- Linehan, M. M. (1993a). *Cognitive-behavioral treatment of borderline personality disorder*. New York, NY: Guilford Press.
- Linehan, M. M. (1993b). *Skills training manual for treating borderline personality disorder*. New York, NY: Guilford Press.
- *Linehan, M. M., Armstrong, H. E., Suarez, A., Allmon, D., & Heard, H. L. (1991). Cognitive-behavioral treatment of chronically parasuicidal borderline patients. *Archives of General Psychiatry, 48*, 1060–1064.
- Linehan, M. M., & Comtois, K. A. (1996). *Lifetime parasuicide count*. Unpublished manuscript, Department of Psychology, University of Washington, Seattle.
- Linehan, M. M., Comtois, K. A., Brown, M. Z., Heard, H. L., & Wagner, A. (2006). Suicide Attempt Self-Injury Interview (SASII): Development, reliability, and validity of a scale to assess suicide attempts and intentional self-injury. *Psychological Assessment, 18*, 303–312.
- *Linehan, M. M., Comtois, K. A., Murray, A. M., Brown, M. Z., Gallop, R. J., Heard, H. L., . . . Lindenboim, N. (2006). Two-year randomized controlled trial and follow-up of dialectical behavior therapy vs therapy by experts for suicidal behaviors and borderline personality disorder. *Archives of General Psychiatry, 63*, 757–766.
- *Linehan, M. M., Dimeff, L. A., Reynolds, S. K., Comtois, K. A., Shaw-Welch, S., Heagerty, P., & Kivlahan, D. R. (2002). Dialectical behavior

- therapy versus comprehensive validation therapy plus 12-step for the treatment of opioid dependent women meeting criteria for borderline personality disorder. *Drug and Alcohol Dependence*, 67, 13–26.
- *Linehan, M. M., Heard, H. L., & Armstrong, H. E. (1993). Naturalistic follow-up of a behavioral treatment for chronically parasuicidal borderline patients. *Archives of General Psychiatry*, 50, 971–974.
- *Linehan, M. M., McDavid, J. D., Brown, M. Z., Sayrs, J. H. R., & Gallop, R. J. (2008). Olanzapine plus dialectical behavior therapy for women with high irritability who meet criteria for borderline personality disorder: A double-blind, placebo-controlled pilot study. *Journal of Clinical Psychiatry*, 69, 999–1005.
- *Linehan, M. M., Schmidt, H., III, Dimeff, L. A., Craft, J. C., Kanter, J., & Comtois, K. A. (1999). Dialectical behavior therapy for patients with borderline personality disorder and drug-dependence. *American Journal on Addictions*, 8, 279–292.
- *Linehan, M. M., Tutek, D. A., Heard, H. L., & Armstrong, H. E. (1994). Interpersonal outcome of cognitive behavioral treatment for chronically suicidal borderline patients. *American Journal of Psychiatry*, 151, 1771–1776.
- MacLehose, R., Reeves, B. C., Harvey, I. M., Sheldon, T. A., Russell, I. T., & Black, A. M. (2000). A systematic review of comparisons of effect sizes derived from randomised and nonrandomised studies. *Health Technology Assessment*, 4, 1–154.
- Matt, G. E., & Cook, T. D. (1994). Threats to the validity of research syntheses. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 503–520). New York, NY: Russell Sage Foundation.
- *McMain, S. F., Links, P. S., Gnam, W. H., Guimond, T., Cardish, R. J., Kormann, L., & Streiner, D. L. (2009). A randomized trial of dialectical behavior therapy versus general psychiatric management for borderline personality disorder. *American Journal of Psychiatry*, 166, 1365–1374.
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & the PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Open Medicine*, 3, 123–130.
- Moher, D., Pham, B., Jones, A., Cook, D. J., Jadad, A. R., Moher, M., Tugwell, P., & Klassen, T. P. (1998). Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*, 352, 609–613.
- Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7, 105–125.
- Oldham, J. M. (2006). Borderline personality disorder and suicidality. *American Journal of Psychiatry*, 163, 20–26.
- Öst, L.-G. (2008). Efficacy of the third wave of behavioral therapies: A systematic review and meta-analysis. *Behaviour Research and Therapy*, 46, 296–321.
- *Prendergast, N., & McCausland, J. (2007). Dialectical behaviour therapy: A 12-month collaborative program in a local community setting. *Behaviour Change*, 24, 25–35.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S., Bryk, A., Cheong, Y. F., Congdon, R. J., & du Toit, M. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638–641.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.
- Rosenthal, R. (1994). Parameter measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York, NY: Russell Sage Foundation.
- Rüsch, N., Schiel, S., Corrigan, P. W., Leihener, F., Jacob, G. A., Olschewski, M., . . . Bohus, M. (2008). Predictors of dropout from inpatient dialectical behavior therapy among women with borderline personality disorder. *Journal of Behavior Therapy and Experimental Psychiatry*, 39, 497–503.
- Rustenbach, J. S. (2003). *Metaanalyse. Eine anwendungsorientierte Einführung* [Meta-Analysis: A user-oriented introduction]. Bern, Germany: Huber.
- Seligman, M. E. P. (1995). The effectiveness of psychotherapy. The *Consumer Reports* study. *American Psychologist*, 50, 956–974.
- Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York, NY: Russell Sage Foundation.
- Shearin, E. N., & Linehan, M. M. (1992). Patient–therapist ratings and relationship to progress in dialectical behavior therapy for borderline personality disorder. *Behavior Therapy*, 23, 730–741.
- Sica, G. T. (2006). Bias in research studies. *Radiology*, 238, 780–789.
- *Simpson, E. B., Yen, S., Costello, E., Rosen, K., Begin, A., Pistorello, J., & Pearlstein, T. (2004). Combined dialectical behavior therapy and fluoxetine in the treatment of borderline personality disorder. *Journal of Clinical Psychiatry*, 65, 379–385.
- Strupp, H. H., Horowitz, L. M., & Lambert, M. J. (Eds.). (1997). *Measuring patient changes in mood, anxiety, and personality disorders: Toward a core battery*. Washington, DC: American Psychological Association.
- Swenson, C. R., Sanderson, C., Dulit, R., & Linehan, M. M. (2001). The application of dialectical behavior therapy for patients with borderline personality disorder on inpatient units. *Psychiatric Quarterly*, 72, 307–324.
- Turner, R. M. (2000). Naturalistic evaluation of dialectical behavior therapy–oriented treatment for borderline personality disorder. *Cognitive and Behavioral Practice*, 7, 413–419.
- *van den Bosch, L. M. C., Koeter, M. W. J., Stijnen, T., Verheul, R., & van den Brink, W. (2005). Sustained efficacy of dialectical behaviour therapy for borderline personality disorder. *Behaviour Research and Therapy*, 43, 1231–1241.
- *van den Bosch, L. M. C., Verheul, R., Schippers, G. M., & van den Brink, W. (2002). Dialectical behavior therapy of borderline patients with and without substance use problems. Implementation and long-term effects. *Addictive Behaviors*, 27, 911–923.
- *van den Bosch, L. M. C., Verheul, R., & van den Brink, W. (2001). Substance abuse in borderline personality disorder: Clinical and etiological correlates. *Journal of Personality Disorders*, 15, 416–424.
- *Verheul, R., van den Bosch, L. M. C., Koeter, M. W. J., de Ridder, M. A. J., Stijnen, T., & van den Brink, W. (2003). Dialectical behaviour therapy for women with borderline personality disorder: 12-month, randomised clinical trial in the Netherlands. *British Journal of Psychiatry*, 182, 135–140.
- Westen, D., Novotny, C. M., & Thompson-Brenner, H. (2004). The empirical status of empirically supported psychotherapies: Assumptions, findings, and reporting in controlled clinical trials. *Psychological Bulletin*, 130, 631–663.
- Zanarini, M. C., Frankenburg, F. R., Dubo, E. D., Sickel, A. E., Trikha, A., Levin, A., & Reynolds, V. (1998). Axis I comorbidity of borderline personality disorder. *American Journal of Psychiatry*, 155, 1733–1739.
- Zimmerman, M., & Mattia, J. I. (1999). Is posttraumatic stress disorder underdiagnosed in routine clinical settings? *Journal of Nervous and Mental Disease*, 187, 420–428.

Appendix

Formulae

Between-Groups Effect Size

$$g_{\text{Hedges}} = \frac{x_{\text{post1}} - x_{\text{post2}}}{\sigma_{\text{pooled}}}$$

where

$$\sigma_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)\sigma_{\text{post1}}^2 + (n_2 - 1)\sigma_{\text{post2}}^2}{n_1 + n_2 - 2}}$$

$$d = g_{\text{Hedges}} \left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1} \right),$$

$$\sigma_d^2 = \frac{n_1 + n_2}{n_1 n_2} + \frac{d_j^2}{2(n_1 + n_2)},$$

where

- σ = standard deviation,
- n = sample size,
- x = intervention outcome score,
- post = postintervention, and
- pre = preintervention.

Within-Group Effect Size

$$d_w = \frac{x_{\text{pre}} - x_{\text{post}}}{\sigma_{\text{Diff}}}$$

where

$$\sigma_{\text{Diff}} = \sqrt{\sigma_{\text{pre}}^2 + \sigma_{\text{post}}^2 - 2r_{\text{pre,post}}\sigma_{\text{pre}}\sigma_{\text{post}}}$$

$$d = d_w \left(1 - \frac{3}{4(n_{\text{pre}} - 1) - 1} \right),$$

$$\sigma_d^2 = \frac{1}{n} + \frac{d_j^2}{2(n - 1)}.$$

Transforming t Statistics or Single-Group Repeated Measures Analyses of Variance Into Hedges's g

$$g_{\text{Hedges}} = t \sqrt{\frac{n_1 + n_2}{n_1 n_2}},$$

respectively

$$d_w = \frac{t}{\sqrt{n}}, \quad t = \sqrt{F}.$$

Transforming Odds Ratios Into Hedges's g

$$g_{\text{Hedges}} = \frac{\sqrt{3}}{\pi \ln(o)},$$

where

$$o = \frac{n_{+I}n_{-C}}{n_{+C}n_{-I}},$$

where

- n_{+I} = number of positive outcome in treatment group,
- n_{-C} = number of negative outcome in control condition,
- n_{+C} = number of positive outcome in control condition, and
- n_{-I} = number of negative outcome in treatment group.

Transforming Product-Moment Correlation and Chi-Square Statistics Into Hedges's g

$$g_{\text{Hedges}} = \frac{r}{\sqrt{1 - r^2}} \sqrt{\frac{(n_1 + n_2 - 2)(n_1 + n_2)}{n_1 n_2}},$$

$$r \approx \Phi = \sqrt{\frac{\chi^2}{N}}.$$

The Empirical Bayes Estimator, δ'_j , of Each Study's Effect

$$\delta'_j = \lambda_j d_j + (1 - \lambda_j) \tilde{y}_0,$$

where

- $\lambda_j = \frac{\tau}{\tau + \sigma_{dj}^2}$ = reliability of each study's effect,
- τ_j = maximum likelihood estimate of the parameter variance,
- $\tilde{y}_0 = \frac{\sum \Delta_j^{-1} d_j}{\sum \Delta_j^{-1}}$ = weighted least square estimator of the grand mean,

where

Δ_j^{-1} = precision of d_j ,

where

$$\Delta_j = \tau_j + \sigma_{dj}^2.$$

The H Statistic

$$H = \sum (\sigma_j^2)^{-1} (d_j - \tilde{y}_0)^2.$$

(Appendix continues)

Hierarchical Linear Model**Unconditional model.***Level 1: Within studies.*

$$\tilde{y}_{jk} = \delta'_{jk} + e_{jk},$$

Level 2: Between outcome measures.

$$\delta'_{jk} = \beta_{0k} + r_{jk},$$

Level 3: Between studies.

$$\beta_{0k} = \gamma_{00} + u_{0k},$$

Conditional model.*Level 1: Within studies.*

$$\tilde{y}_{jk} = \delta'_{jk} + e_{jk},$$

Level 2: Between outcome measures.

$$\delta'_{jk} = \beta_{0k} + r_{jk},$$

Level 3: Between studies.

$$\beta_{0k} = \gamma_{00} + \sum_{p=1} \gamma_p W_{pk} + u_{0k},$$

where

 W = study characteristic used as a predictor for the study effect.Outcome measures $j = 1, \dots, J$ within studies $k = 1, \dots, K$

$$e_{jk} \sim N(0; \sigma_{jk}^2)$$

$$r_{jk} \sim N(0; \tau_{\delta'})$$

$$u_{0k} \sim N(0; \tau_{\beta})$$

Received December 10, 2009

Revision received July 27, 2010

Accepted July 28, 2010 ■