Available online at www.sciencedirect.com
ScienceDirect

Behaviour Research and Therapy 46 (2008) 296–321

BEHAVIOUR RESEARCH AND THERAPY

www.elsevier.com/locate/brat

Efficacy of the third wave of behavioral therapies: A systematic review and meta-analysis

Lars-Göran Öst*

Department of Psychology, Stockholm University, S-106 91 Stockholm, Sweden

Received 17 September 2007; received in revised form 11 December 2007; accepted 14 December 2007

Abstract

During the last two decades a number of therapies, under the name of the third wave of cognitive behavior therapy (CBT), have been developed: acceptance and commitment therapy (ACT), dialectical behavior therapy (DBT), cognitive behavioral analysis system of psychotherapy (CBASP), functional analytic psychotherapy (FAP), and integrative behavioral couple therapy (IBCT). The purposes of this review article of third wave treatment RCTs were: (1) to describe and review them methodologically, (2) to meta-analytically assess their efficacy, and (3) to evaluate if they currently fulfil the criteria for empirically supported treatments. There are 13 RCTs both in ACT and DBT, 1 in CBASP, 2 in IBCT, and none in FAP. The conclusions that can be drawn are that the third wave treatment RCTs used a research methodology that was significantly less stringent than CBT studies; that the mean effect size was moderate for both ACT and DBT, and that none of the third wave therapies fulfilled the criteria for empirically supported treatments. The article ends with suggestions on how to improve future RCTs to increase the possibility of them becoming empirically supported treatments.
© 2008 Elsevier Ltd. All rights reserved.

Keywords: Third wave of CBT; ACT; DBT; CBASP; IBCT; Systematic review; Meta-analysis

Introduction

Behavior therapy (BT) was developed in the 1950s more or less independently in three countries: (1) South Africa, where Joseph Wolpe developed a treatment, systematic desensitization, for phobias and other anxiety disorders and published the book *Psychotherapy by reciprocal inhibition* in 1958; (2) USA, where Ogden Lindsley used operant techniques in working with schizophrenic patients; and (3) England, where Hans Eysenck was instrumental in developing an alternative to psychoanalysis, which he in his famous 1952 review article found to be no more effective than no treatment.

Thus, BT is considered to be the first ‘‘wave’’ of a scientifically based psychotherapy. The second wave is cognitive therapy (CT), developed by Aaron Beck in the 1970s with its first application to depression, and later to anxiety disorders and eating disorders (Hayes, 2004). In the late 80s–early 90s, there was a merge between BT and CT into what has been named cognitive behavior therapy (CBT), and this is the form of therapy having the largest evidence base today (Roth & Fonagy, 2005).

*Tel.: +46 8 163821; fax: +46 8 161002.
E-mail address: ost@psychology.su.se

0005-7967/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.brat.2007.12.005

During the last 10–15 years a number of new treatments, or extensions from previous CBT treatments, have appeared on the psychotherapy arena. Hayes (2004) described the so-called third wave of BT:

> *Grounded in an empirical, principle-focused approach, the third wave of behavioral and cognitive therapy is particularly sensitive to the context and functions of psychological phenomena, not just their form, and thus tends to emphasize contextual and experiential change strategies in addition to more direct and didactive ones. These treatments tend to seek the construction of broad, flexible, and effective repertoires over an eliminative approach to narrowly defined problems, and to emphasize the relevance of the issues they examine for clinicians as well as clients. The third wave reformulates and synthesizes previous generations of behavioral and cognitive therapy and carries them forward into questions, issues, and domains previously addressed primarily by other traditions, in hope of improving both understanding and outcomes.* (Hayes, 2004, p. 658; italics in original).

The third wave therapies share some features, but there are also differences. Among the common characteristics are a focus on mindfulness, acceptance, defusion, the patient's values in life, relationships, the rationale for how the treatment works, and the client–therapist relationship (e.g. Hayes, 2004). However, other authors (e.g. Hofmann & Asmundson, 2008) have argued that some of these characteristics have been part of CBT for a long time and question if it is correct to talk about a third wave. Furthermore, the treatments differ concerning their theoretical basis, where some of the therapies are based on behavioral analysis or radical behaviorism, while others do not subscribe to this.

For this review, treatments that were developed to be a primary treatment for psychiatric disorders and have been described as a third wave therapy (e.g. Hayes, 2004) will be included: (1) acceptance and commitment therapy (ACT; Hayes, Strosahl, & Wilson, 1999), (2) dialectical behavior therapy (DBT; Linehan, 1993), (3) cognitive behavioral analysis system of psychotherapy (CBASP; McCullough, 2000), (4) functional analytic psychotherapy (FAP; Kohlenberg & Tsai, 1991), and (5) integrative behavioral couple therapy (IBCT; Jacobson & Christensen, 1996). Thus, the following treatments have not been included: mindfulness-based stress reduction (Kabat-Zinn, 1990), which mainly has been tested in RCTs for stress associated with various somatic disorders (see the review by Bishop (2002), and meta-analysis by Grossman, Niemann, Schmidt, & Walach, 2004), mindfulness-based cognitive therapy (Segal, Williams, & Teasdale, 2002), which so far primarily has been used as a prevention method in depression, and metacognitive therapy (Wells, 2000), which the originator sees as a form of traditional CBT (Wells, 2006). For a more detailed description of the three waves, see Hayes (2004).

The third wave therapies have attracted a great deal of interest, but also some critique. For example, Corrigan (2001) argued that these therapies were "getting ahead of the data" (p. 192) and questioned if their proponents were committed to the principles of empirical validation laid down by the first wave behavior therapists. He further characterized people interested in these therapies as "devotees of interventions that lack the data to support them" (p. 192). His arguments were rebutted by Hayes (2002), and in a later paper Hayes, Masuda, Bissett, Luoma, and Guerrero (2004) reviewed the empirical database. However, so far there has not been an independent review of the evidence base when it comes to treatment outcome studies done on the third wave therapies.

The purposes of this article are: (1) to describe the third wave RCTs and review them from a methodological point of view, (2) to review the efficacy of third wave treatments using meta-analysis procedure, and (3) to conduct a preliminary evaluation of the third wave treatments in relation to the criteria for empirically supported treatments (ESTs) set forth by the APA Division 12 Task Force in 1995 (Chambless, Baker, Baucom, Beutler, & Calhoun, 1998).

Method

Literature search

PsycINFO and Medline were searched from 1985 to mid 2007 with the following search words:

- Acceptance and Commitment Therapy or ACT
- Dialectical Behavior Therapy or DBT

- Cognitive Behavioral Analysis System of Psychotherapy or CBASP
- Functional Analytic Psychotherapy or FAP
- Integrative Behavioral Couple Therapy or IBCT

All abstracts were read, and when there was an indication of a group of patients receiving the particular treatment being compared with another group in a randomized clinical trial (RCT) the entire article was retrieved. Studies using single case designs were excluded since there is no consensus yet regarding the calculation of effect sizes. The reference lists in the retrieved articles were then checked against the database search and any other articles that might fulfil the inclusion criteria were retrieved.

Inclusion criteria

In order to be included in the review and meta-analysis, the following criteria were to be adhered to:

- a study had to be published, or in press, in an English language journal;
- participants had to be randomly allocated to either treatment and control, or to two or more active treatments;
- one of the third wave treatments had to be used.

Methodological stringency

Since the criteria for EST specify that the studies on which to judge a treatment must have used rigorous methodology (Chambless & Ollendick, 2001), a rating of the methodological stringency or quality of the third wave studies is warranted. This can be informative and of value in itself, and also be used as a moderator variable in a meta-analysis.

Construction of a rating scale

Foa and Meadows (1997) described a "golden standard" for therapy outcome studies in PTSD. This was later used as the basis for a rating scale by Tolin (1999) containing 13 items rated on a 0–2 scale, with each step described and usually exemplified. He used it for a revised meta-analysis on PTSD. For the present study Tolin's scale was modified to be suitable for outcome studies in general: one item was deleted and 10 items were added using the same type of descriptions and examples.

The psychotherapy outcome study methodology rating scale

The scale consists of the following 22 items (see Appendix A): (1) clarity of sample description, (2) severity/chronicity of the disorder, (3) representativeness of the sample, (4) reliability of the diagnosis in question, (5) specificity of outcome measures, (6) reliability and validity of outcome measures, (7) use of blind evaluators, (8) assessor training, (9) assignment to treatment, (10) design, (11) power analysis, (12) assessment points, (13) manualized, replicable, specific treatment programs, (14) number of therapists, (15) therapist training/experience, (16) checks for treatment adherence, (17) checks for therapist competence, (18) control of concomitant treatments, (19) handling of attrition, (20) statistical analyses and presentation of results, (21) clinical significance, (22) equality of therapy hours (for non-WLC designs only). Each item is rated as 0 = poor, 1 = fair, and 2 = good, and each step has a verbal description.

Psychometric data

The internal consistency of the scale was good, with a Cronbach's α of 0.86. In order to assess the inter-rater reliability of the scale, an advanced graduate student in clinical psychology received 2 h of training in the use of the scale by the author, with various outcome studies as training examples. Then the student blindly rated a random selection of 20% of the studies and the ratings were compared with those of the author.

The intra-class correlation for the total score was 0.92, and the *kappa* coefficients on the individual items varied between 0.50 and 1.00, with a mean of 0.75, indicating a good inter-rater reliability.

Comparison with CBT studies

Since the third wave therapies have developed from traditional CBT, and this is by far the form of psychotherapy with most RCTs and ESTs (Roth & Fonagy, 2005), it seems logical to make a comparison between RCTs of the third wave therapies and those of traditional CBT (henceforth CBT) regarding methodological stringency. In order to assess whether the obtained methodology score is equal to or different from CBT studies, a "comparison" sample of studies was collected. For each of the third wave studies a "twin" CBT study published in the same journal the same, or ±1, year was retrieved. The reason for this time window is that the psychotherapy research methodology is continuously developing and it would be unfair to compare old studies of one wave with new studies of another. It was possible to obtain directly matching CBT studies for 18 of the 29 third wave studies. For the remaining 11 studies, which were published in journals rarely used by traditional CBT researchers, a comparison study was obtained from the three main outlets for CBT outcome research; *Journal of Consulting and Clinical Psychology*, *Behavior Therapy*, and *Behaviour Research and Therapy*, with 3–4 studies from each journal.

The selection of comparison studies was done in the following way. First, all RCTs of the volume in question were listed and numbered according to when it was published (the page number). Second, studies by me or any of my co-workers were deleted to avoid biased ratings. Third, studies on children and adolescents were deleted since the third wave studies only have adult or geriatric samples. Fourth, a computerized randomization procedure (www.randomizer.org) was used to select which studies to include in the comparison sample. The traditional CBT studies used for this comparison are listed in Appendix B.

Meta-analysis

In order to avoid dependence among multiple effect sizes from the same study, a decision was made as to which was the main outcome measure for each study, and this was used in the meta-analysis. The controlled ES was calculated post-treatment by dividing the difference between the treatment mean and the control (comparison) mean with the pooled standard deviation of the two conditions.

The uncontrolled ES (within-group) was calculated by dividing the mean change from pre to post with the pre-treatment SD, and the mean change from pre to follow-up with the pre-treatment SD (Feske & Chambless, 1995; Morris & DeShon, 2002).

The meta-analysis was performed using the comprehensive meta-analysis, version 2 software (Biostat, Inc., 2006), weighting the ESs by the reciprocal of the sampling variance (taking sample size into consideration) and correcting for small samples by calculating Hedges' g. Homogeneity among ESs was assessed with the Q-statistic and a significant heterogeneity was followed up with moderator analysis. A random model was used when a significant Q-value was obtained (Lipsey & Wilson, 2001).

Criteria for ESTs

The APA Division 12 Task Force (Chambless et al., 1998) developed the criteria for empirically supported therapies. In order for a psychological treatment to be considered empirically supported (well-established), the following criteria have to be fulfilled:

I. At least two good between-group design experiments must demonstrate efficacy in one or more of the following ways:
   A. Superiority to pill or psychotherapy placebo, or to other treatment
   B. Equivalence to already established treatment with adequate sample sizes[1] (or)

[1] In the original Task Force report this was defined as "about 30 per group".

Table 1
Means (SDs) and *t*-values for different background and therapy variables

Variable	Third wave ($n = 29$)	CBT ($n = 29$)	*t*-value
1. Number of participants starting therapy	70.8 (121.6)	68.0 (25.7)	0.12
2. Attrition (percent of those starting)	20.6 (15.1)	14.3 (8.4)	1.94
3. Number of completers	54.6 (92.8)	56.3 (20.7)	0.10
4. Cell size (completers/number of conditions)	23.2 (31.2)	22.4 (9.9)	0.13
5. Proportion of women	77.4 (23.8)	68.2 (24.4)	1.45
6. Mean age of the sample	39.3 (9.6)	38.2 (8.4)	0.48
7. Number of therapy weeks	21.8 (18.0)	12.3 (6.6)	2.62*
8. Number of therapy sessions	19.3 (18.3)	12.6 (5.7)	1.72
9. Number of therapy hours	30.0 (29.8)	16.0 (9.0)	2.33*
10. Follow-up (months since post-assessment)	4.4 (5.5)	9.4 (12.2)	2.01*

*$p < 0.05$.

II. A large series[2] of single-case design experiments must demonstrate efficacy with
   A. use of good experimental design, and
   B. comparison of intervention to another treatment.
III. Experiments must be conducted with treatment manuals or equivalent clear description of treatment.
IV. Characteristics of samples must be specified.
V. Effects must be demonstrated by at least two investigators or teams.

Criteria I and III–V will be applied in the present review to judge the empirical status of the third wave therapies. Of course, it is necessary to have a work group (e.g. a Task Force) consisting of different experts to make a final judgement on this issue. However, a fellow of the research community must be allowed to offer his/her preliminary evaluation.

Results

Description of the third wave studies

The third wave and the traditional CBT studies are compared on a number of background and therapy variables in Table 1. The significant differences obtained showed that the third wave studies had longer therapies and higher number of therapy hours, whereas CBT studies had longer follow-up periods. A closer look at the different third wave studies is warranted.

ACT studies

There are 13 RCTs with a total of 677 participants, in which ACT, singly or in combination with another treatment, has been compared to a control group, or another active treatment (see Table 2). The disorders focused on in these studies vary to a large extent, with two on depression, two on psychotic symptoms, and two on stress. The remaining seven studies focus on different disorders. Six of the studies did not use a diagnostic system (DSM, ICD, or a similar standardized system) to diagnose the participants, which is remarkable since the earliest study was published in 1986 and DSM-III (APA, 1980) should have been available when it was carried out. There are large variations in the number of participants starting the study (18–124), the attrition rate (0–37%), and the mean cell size of completers (6–39). Three of the studies had only female subjects and the mean proportion was 68%. All studies had middle-aged participants with average age ranging from 30.5 to 50.9, with a mean of 39.4. The treatment varied from 1 to 16 weeks, from 1 to 48 sessions, and from 3 to 24 h. Three studies had no follow-up at all, and only two reported a 1-year follow-up, with a mean of 4.2 months.

[2] In the original Task Force report this was defined as at least 9.

Table 2
Characteristics of the ACT studies

Study	Disorder	Diagnostic system	Comparison[a]	N at start	Attrition[b]	Cell size[c]	% fem.	M age	Tx weeks	Tx sess.	Tx hours[d]	F-up months
Zettle (1986)	Depression	NI	CT	18	NI	6	100	NI	12	12	12.0	3
Zettle (1989)	Depression	NI	CT	37	16.2	10	100	41.3	12	12	10.8	2
Bond (2000)	Stress	NI	IPP	90	27.8	22	50	36.4	14	3	9.8	3
Bach (2002)	Psychotic symptoms	DSM-IV	TAU	80	12.5	35	36	39.4	2	4	4.0	4
Zettle (2003)	Mathematics anxiety	NI	SD	33	27.8	12	81	30.5	6	6	6.0	0
Dahl (2004)	Stress and pain	NI	MTAU	19	0.0	10	81	40.0	4	4	4.0	6
Gifford (2004)	Smoking	NI	NRT	76	35.5	25	59	43.0	7	14	16.3	12
Hayes (2004)	Opiate dependence	DSM-III-R	ITFP	124	37.1	26	51	42.2	16	48	24.4	5
Gaudiano (2006)	Psychotic symptoms	DSM-IV	ETAU	40	5.0	19	36	40.0	3	4	3.0	4
Woods (2006)	Trichotillomania	DSM-IV	WLC	28	10.9	13	89	35.0	12	10	12.0	3
Gratz (2006)	Borderline PD	DSM-IV	TAU	24	8.3	11	100	33.2	14	14	21.0	0
Lundgren (2006)	Epilepsy	EEG	ST	27	0.0	14	52	40.7	4	4	9.0	12
Gregg (2007)	Diabetes	Blood test	EA	81	3.7	39	47	50.9	1	1	7.0	0
Mean				52.1	15.4	18.5	67.9	39.4	8.2	10.5	10.7	4.2
SD				33.9	13.4	10.2	24.6	5.3	5.2	12.1	6.6	4.0

[a] CT = cognitive therapy, IPP = Innovation Promotion Program, TAU = treatment as usual, MTAU = medical treatment as usual, ETAU = enhanced treatment as usual, SD = systematic desensitization, NRT = nicotine replacement treatment, ITFP = Intensive Twelve Step Facilitation Program, WLC = waiting list control, ST = supportive therapy, EA = education alone.

[b] Percent dropouts of the number who started treatment.

[c] Number completing treatment divided by the number of conditions in the study and rounded to the nearest integer.

[d] Number of sessions × session length. When mean number of sessions is given, that figure is used instead of the maximum number in calculating treatment hours. NI = no information is given in the article.

DBT studies

There are also 13 RCTs on DBT with a total of 539 participants (see Table 3). However, in two of these (Simpson, Yen, Costello, Rosen, & Begin, 2004; Soler, Pasqual, Campins, Barrachina, & Puigdemont, 2005) all patients received DBT together with placebo or an active drug, whereas in two studies (Lynch, Morse, Mendelson, & Robins, 2003, 2007) DBT plus antidepressant medication (ADM) was compared with ADM alone. Since DBT was developed for borderline personality disorder (BDP), it is understandable that nine of the studies focused on BPD patients, while two were on eating disorders and two on depression in older patients. In all studies different versions of the DSM were used to diagnose the participants. The number of subjects starting the studies varied from 23 to 101, the attrition rate from 6% to 59%, and the mean cell size of completers from 7 to 30. Altogether nine of the studies had exclusively female samples and the mean across the DBT studies was 92%. The participants' mean age varied from 22.5 to 66.0, with an average of 38.5. The original DBT treatment for BPD lasted for 1 year, and five of the studies had that length. However, one had 26 weeks, and two only had 12 weeks of therapy. The eating disorder studies both used 20 weeks and the depression studies had 28 and 24 weeks, respectively. Number of therapy hours varied from 17 to 138. Six of the studies had no follow-up at all, whereas two had 1 year, with the mean being 4 months.

CBASP study

There is only one CBASP study (Table 4) published so far (Keller, McCullough, Klein, Arnow, & Dunner, 2000). This is a multi-site study with 681 chronically depressed patients starting therapy and 23.8% dropping

Table 3
Characteristics of the DBT studies

Study	Disorder	Diagnostic system	Comparison[a]	*N* at start	Attrition[b]	Cell size[c]	% fem.	*M* age	Tx weeks	Tx sess.	Tx hours[d]	F-up months
Linehan (1991)	Borderline PD	DSM-III	TAU	46	37.0	15	100	NI	52	NI	NI	12
Linehan (1999)	Borderline PD	DSM-III-R	TAU	28	35.7	9	100	30.4	52	NI	43.1	4
Linehan (2002)	Borderline PD	DSM-IV	CVT	23	13.0	10	100	36.1	52	33.2	46.5	4
Linehan (2006)	Borderline PD	DSM-IV	CTBE	101	41.6	30	100	29.3	52	80.5	137.5	12
Turner (2000)	Borderline PD	DSM-III	CCT	24	41.7	7	79	22.5	52	NI	66.5	0
Koons (2001)	Borderline PD	DSM-III-R	TAU	28	28.6	10	100	35.0	26	NI	50.9	0
Verheul (2003)	Borderline PD	DSM-IV	TAU	58	58.6	12	100	34.9	52	NI	NI	6
Simpson (2004)	Borderline PD	DSM-IV	Fluox.	25	20.0	10	100	35.3	12	25	36.0	0
Soler (2005)	Borderline PD	DSM-IV	Olanz.	60	30.0	21	87	27.0	12	12	30.0	0
Safer (2001)	Bulimia nervosa	DSM-IV	WLC	31	9.7	14	100	34.0	20	20	16.7	0
Telch (2001)	Binge eating	DSM-IV	WLC	44	22.7	17	100	50.0	20	20	40.0	6
Lynch (2003)	Depression	DSM-III-R	ADM	36	5.6	17	85	66.0	28	28	70.0	0
Lynch (2007)	Depression + PD	DSM-IV	ADM	35	8.6	16	46	61.4	24	48	72.0	6
Mean				41.5	27.1	14.4	92.1	38.5	34.9	25.5	55.4	3.9
SD				21.7	15.7	6.0	15.6	13.6	17.1	12.3	32.2	4.4

[a] TAU = treatment as usual, CVT = comprehensive validation therapy, CTBE = community treatment by experts, CCT = client centered therapy, Fluox. = fluoxetine, Olanz. = olanzapine, WLC = waiting list control, ADM = antidepressant medication.

[b] Percent dropouts of the number who started treatment.

[c] Number completing treatment divided by the number of conditions in the study and rounded to the nearest integer.

[d] Number of sessions × session length. When mean number of sessions is given, that figure is used instead of the maximum number in calculating treatment hours. NI = no information is given in the article.

Table 4
Characteristics of the CBASP, and IBCT studies

Study	Disorder	Diagnostic system	Comparison[a]	*N* at start	Attrition[b]	Cell size[c]	% fem.	*M* age	Tx weeks	Tx sess.	Tx hours[d]	F-up months
Keller (2000)	Depression	DSM-IV	ADM	681	23.8	173	65	43.0	12	16.1	16.1	0
Jacobson (2000)	Marital discord	NA	TBCT	21	9.5	10	50	41.3	NI	21.0	21.0	0
Christensen (2004)	Marital discord	NA	TBCT	134	6.0	63	50	42.6	36	23.5	23.5	24
Mean				279	13.1	81.8	55.0	42.3	24.0	20.2	20.2	8.0
SD				352	9.4	83.4	8.7	0.9	17.0	3.8	3.8	13.9

[a] ADM = antidepressant medication, TBCT = traditional behavioral couple therapy.

[b] Percent dropouts of the number who started treatment.

[c] Number completing treatment divided by the number of conditions in the study and rounded to the nearest integer.

[d] Number of sessions × session length. When mean number of sessions is given, that figure is used instead of the maximum number in calculating treatment hours. NA = not applicable. NI = no information is given in the article.

out. It still leaves a big cell size (173) of mainly female (65%), middle-aged ($M = 43$ years) patients. They received 16 sessions of therapy across a 12-week period. No follow-up was reported.

IBCT studies

There are only two published IBCT studies on marital discord with a total of 155 participants, both from the collaboration between Neil Jacobson (Seattle) and Andrew Christensen (Los Angeles). The first has a small cell size, which is understandable for a pilot study, but the second one has a very respectable cell size. The gender proportion is even since heterosexual couples were the participants. They received a mean of 23.5 sessions, which is normal for this treatment.

FAP study

The literature search failed to find any RCT on FAP. The closest to one was a study by Kohlenberg, Kanter, Bolling, Parker, and Tsai (2002) on depressed patients. During one year the therapists used standard CT and for the next year they used FAP enhanced CT. However, a clear-cut conclusion cannot be drawn from this study since the participants were not randomized to conditions.

Comparison with CBT studies on descriptive variables

The ACT and DBT studies were compared to the randomized sample of CBT studies on the descriptive variables (Table 5). The proportion of studies using any diagnostic system was significantly (Fisher's exact probability test, $P_2 = 0.0002$) lower in the ACT studies (54%) than in the DBT (100%) and CBT studies (100%).

For the continuous variables one-way ANOVAs were used followed by Tukey's test if the *F*-value was significant. The number of starting and of completing participants was lower in DBT studies than in CBT studies, whereas the ACT studies did not differ from either in this respect. The attrition rate was higher in DBT studies than in both ACT and CBT studies. The completer cell size was higher in CBT studies than in DBT studies, whereas ACT studies were in between the other two. DBT studies had a higher proportion of females than both ACT and CBT studies. The number of therapy weeks was lower in ACT than in DBT studies, whereas CBT studies did not differ from either, and the number of therapy hours obtained was higher in DBT studies than in both ACT and CBT studies. There was no significant difference regarding mean age, number of therapy sessions, and length of follow-up.

Methodological stringency

When applying the methodology rating scale described above, the total mean score for the 32 third wave studies was 19.6 (SD = 4.9), which was significantly lower ($t(56) = 6.82$, $p < 0.0001$) than the mean of the CBT

Table 5
Means (SDs) and *F*-values for different background variables

Variable	ACT ($n = 13$)	DBT ($n = 13$)	CBT ($n = 26$)	*F*-value
1. Number of participants starting therapy	52.1 (33.9)^{ab}	41.5 (21.7)^{a}	76.5 (38.4)^{b}	5.31**
2. Attrition (percent of those starting)	15.4 (13.4)^{a}	27.1 (15.7)^{b}	16.1 (8.0)^{a}	4.52*
3. Number of completers	41.9 (23.2)^{ab}	28.8 (12.1)^{a}	62.1 (32.0)^{b}	7.56***
4. Cell size (completers/number of conditions)	18.5 (10.2)^{ab}	14.4 (6.0)^{a}	24.0 (11.4)^{b}	4.26*
5. Proportion of women	67.9 (24.6)^{b}	92.1 (15.6)^{a}	69.1 (24.4)^{b}	5.21**
6. Mean age of the sample	39.4 (5.3)	38.5 (13.6)	37.8 (8.9)	0.11
7. Number of therapy weeks	8.2 (5.2)^{a}	34.9 (17.1)^{b}	17.2 (29.0)^{ab}	4.81*
8. Number of therapy sessions	10.5 (12.1)	25.5 (12.3)	19.7 (39.1)	1.32
9. Number of therapy hours	10.7 (6.6)^{a}	55.4 (32.2)^{b}	22.0 (32.4)^{a}	8.19***
10. Follow-up (months since post-assessment)	4.2 (4.0)	3.9 (4.4)	9.6 (12.8)	2.18

^{a,b}Means with different superscript differs significantly ($p < 0.05$ or lower). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

studies (27.8, SD = 4.2). The total mean score for ACT was 18.1 (SD = 5.0) and for DBT 19.4 (SD = 3.9), and the ANOVA yielded a significant $F(2, 49) = 30.79$, $p<0.0001$, with both means being lower ($p<0.0001$) than the CBT mean. Since some of the studies could not obtain the maximum score (44) of the scale, e.g. due to their design (comparison with a waiting list group), each study's total score was recalculated to a percentage of the maximum score possible for that study. On this measure the CBT mean (63.6, SD = 9.4) was significantly higher ($t(56) = 6.76$, $p<0.0001$) than the mean for the third wave studies (44.9, SD = 11.5). The mean for ACT was 41.4 (SD = 11.7) and for DBT 44.4 (SD = 8.9), and the ANOVA yielded a significant $F(2, 49) = 31.1$, $p<0.0001$, with both means being lower ($p<0.0001$) than the CBT mean.

For 11 of the 29 third wave studies it was not possible to find a matching CBT study in terms of journal and year of publication. It is therefore important to test if the obtained significant difference favoring CBT studies is caused by the 11 CBT studies published in the three major journals of the field. The comparison regarding the 11 different journal studies showed a significantly higher ($t(20) = 4.06$, $p<0.001$) mean for CBT (28.1, SD = 5.3) than for the third wave studies (18.6, SD = 5.9). However, the comparison between third wave and CBT studies published in the same journals also yielded a significantly higher ($t(34) = 5.49$, $p<0.0001$) mean for CBT (27.6, SD = 3.5) than for the third wave studies (20.1, SD = 4.5). Thus, the overall difference between third wave and CBT studies cannot be caused by the 11 CBT studies in the three major journals being that much better than their third wave counterparts.

In order to get a closer look at the variables that differentiated the three groups of studies, one-way ANOVAs were done, followed by Scheffé's test when significant (see Table 6). Since there are 22 variables, an α of 0.01 was used. Significant differences were obtained for 11 of the 22 variables in the scale. CBT had significantly higher means than both ACT and DBT on items 4—reliability of the diagnosis, 6—reliability and validity of outcome measures, 16—checks for treatment adherence, and 18—control of concomitant treatments. CBT had significantly higher means than ACT only on items 3—representativeness of the sample, 9—assignment to treatments, 14—number of therapists, and 15—therapist training/experience. CBT had significantly higher means than DBT only on items 20—statistical analyses and presentation of data, 21—clinical significance, and 22—equality of therapy hours. CBT and ACT were not significantly different on items 20, 21, and 22, whereas CBT and DBT did not differ significantly on items 3, 9, 14, and 15.

Table 6
Means (SDs) and *F*-values for the different variables in the psychotherapy research methodology scale

Variable	ACT ($n = 13$)	DBT ($n = 13$)	CBT ($n = 26$)	*F*-value
1. Clarity of sample description	1.23 (0.73)	1.54 (0.52)	1.62 (0.50)	2.02
2. Severity/chronicity of the disorder	1.31 (0.86)	1.54 (0.52)	1.73 (0.45)	2.26
3. Representativeness of the sample	1.08 (0.76)[a]	1.46 (0.52)[ab]	1.73 (0.45)[b]	5.99*
4. Reliability of the diagnosis in question	0.15 (0.38)[a]	0.77 (0.60)[b]	1.32 (0.56)[c]	21.17**
5. Specificity of outcome measures	1.77 (0.60)	1.77 (0.44)	2.00 (0.00)	2.56
6. Reliability and validity of outcome measures	1.54 (0.66)[a]	1.23 (0.44)[a]	2.00 (0.00)[b]	18.00**
7. Use of blind evaluators	0.31 (0.48)	0.77 (0.44)	0.58 (0.50)	3.01
8. Assessor training	0.31 (0.63)	0.69 (0.63)	0.77 (0.82)	1.78
9. Assignment to treatment	0.85 (0.38)[a]	1.00 (0.00)[ab]	1.19 (0.40)[b]	4.69*
10. Design	1.23 (0.73)	1.15 (0.80)	1.62 (0.64)	2.40
11. Power analysis	0.00 (0.00)	0.15 (0.56)	0.38 (0.80)	1.71
12. Assessment points	0.92 (0.64)	0.77 (0.73)	1.27 (0.60)	2.99
13. Manualized, replicable, specific treatment programs	1.54 (0.66)	1.54 (0.52)	1.69 (0.55)	0.47
14. Number of therapists	0.23 (0.44)[a]	0.77 (0.44)[b]	1.08 (0.56)[b]	12.21**
15. Therapist training/experience	0.69 (0.75)[a]	0.92 (0.86)[ab]	1.42 (0.64)[b]	4.96*
16. Checks for treatment adherence	0.15 (0.38)[a]	0.31 (0.48)[a]	0.92 (0.80)[b]	7.70**
17. Checks for therapist competence	0.00 (0.00)	0.00 (0.00)	0.23 (0.51)	2.56
18. Control of concomitant treatments	0.23 (0.60)[a]	0.23 (0.44)[a]	1.00 (0.49)[b]	14.94**
19. Handling of attrition	0.85 (0.80)	0.62 (0.77)	1.19 (0.75)	2.65
20. Statistical analyses and presentation of results	1.69 (0.63)[ab]	1.54 (0.52)[a]	2.00 (0.00)[b]	6.36*
21. Clinical significance	0.69 (0.75)[ab]	0.31 (0.63)[a]	1.04 (0.66)[b]	5.15*
22. Equality of therapy hours (for non-WLC designs only)	1.55 (0.82)[a]	0.36 (0.81)[b]	1.86 (0.47)[a]	19.25**

[a,b,c] Means with different superscript differs significantly ($p<.05$ or lower). *$p<0.01$, **$p<0.001$.

ACT had a significantly higher mean than DBT on item 22, whereas DBT was higher than ACT on items 4 and 14.

One possible explanation for the differences observed is that third wave studies have been funded by grants less often than CBT studies, and thus have not had the budget to do what is required for the methodological variables. In order to investigate this possibility, the author notes and acknowledgments of each study were scrutinized for this information. A total of 21 third wave studies and 23 CBT studies reported grant support, a non-significant difference. When ACT, DBT, and CBT studies were compared, the Fisher's exact probability test was significant ($P_2 = 0.02$). This was followed by pairwise comparisons showing that more DBT studies than ACT studies had grants ($P_2 = 0.03$), whereas the proportions for ACT and CBT ($P_2 = 0.22$) and DBT and CBT ($P_2 = 0.99$) did not differ significantly.

Specific methodological issues

In this section the studies will be reviewed regarding some methodological issues that are particularly important when it comes to drawing conclusions from the results obtained in the studies.

Designs

Third wave studies have included studies comparing the active treatment to a waiting list control condition. This is adequate as a first step, but WLC is the weakest possible control. That is why eight studies (4 ACT, 4 DBT) have compared the active treatment to treatment-as-usual (TAU), which, however, means a number of problems. First, it is very difficult to know exactly what this treatment is, and it can change across the period of the study. Second, the researchers have very little knowledge of what goes on in this treatment because the sessions are not subject to audio/video recording to enable adherence and competence ratings. Third, the TAU patients usually get markedly less hours of treatment than patients in active treatments. This is the case in the two ACT studies (Bach & Hayes, 2002; Gratz & Gunderson, 2006) and four DBT studies (Koons, Robins, Tweed, Lynch, & Gonzalez, 2001; Linehan, Armstrong, Suarez, Allmon, & Heard, 1991, 1999; Verheul et al., 2003). Even the well-designed study by Linehan, Comtois, Murray, Brown, and Gallop (2006) had different number of therapy hours for the comparison group, community treatment by experts (CTBE) ($M = 56.0$), which was lower than that for DBT ($M = 137.5$). It is only when the authors counted each group therapy session of $2\frac{1}{2}$ h as 20 min (without giving an explanation) that the mean was similar for DBT (64.7). The difference in amount of therapy is in itself a threat to the internal validity of the study.

Placebo controls are from a methodological standpoint a better alternative than TAU. Bond and Bunce (2000) used what they called Innovation Promotion Program for stressed people in a media company, and Lundgren, Dahl, Melin, and Kies (2006) used supportive therapy for epilepsy. However, a placebo group needs to be described to the patients as a valid treatment, and the credibility (Borkovec & Nau, 1972) of both treatments should be assessed. The latter was not done in these studies.

A treatment method that in previous research has been found effective for the disorder in question is the most stringent comparison condition to use. This was done in five ACT studies. Zettle and Hayes (1986) and Zettle and Rains (1989) compared with CT for depression; Zettle (2003) compared with systematic desensitization for mathematics anxiety; Gifford, Kohlenberg, Hayes, Antonuccio, and Piasecki (2004) compared with nicotine replacement treatment (NRT) in smoking cessation; and Hayes, Wilson, Gifford, Bissett, and Piasecki (2004) compared with intensive twelve step facilitation (ITSF) and methadone maintenance (MM) in opiate addicts. However, none of these studies have assessed treatment credibility. DBT has not yet been compared to another treatment, which in previous research has been found effective for BPD. This is understandable since the alternatives are even more time consuming (e.g. schema-focused therapy and transference-focused psychotherapy, Giesen-Bloo, van Dyck, Spinhoven, van Tilburg, & Dirksen, 2006, which takes 3 years, or psychoanalytically oriented partial hospitalization, Bateman & Fonagy, 1999, taking 18 months). CBASP was compared to ADM (Keller et al., 2000).

Combination of treatments

The third wave therapies are themselves therapeutic packages and so far very few dismantling studies (see Linehan, Dimeff, Reynolds, Comtois, & Welch, 2002, for an exception) have been done, which is understandable at this stage of development. However, some studies have complicated matters by combining ACT or DBT with another active treatment that in previous research has been shown to be effective for the disorder in question. Gratz and Gunderson (2006) combined ACT, DBT, BT and emotion-focused psychotherapy for BPD and called the treatment emotion regulation group intervention. Woods, Wetterneck, and Flessner (2006) combined ACT and habit reversal training for trichotillomania. Lundgren et al. (2006) combined ACT and behavioral seizure control techniques for epilepsy. Turner (2000) combined DBT and psychodynamic techniques for BPD. With the designs used in these studies it is impossible to know what effect ACT or DBT had on the outcomes.

Therapists

In order to avoid a confounding therapist and treatment condition it is necessary that treatment is delivered by more than one therapist. In the ACT studies, Gifford et al. (2004) and Hayes, Wilson, et al. (2004) both used four therapists, whereas Zettle and Rains (1989), Bach and Hayes (2002), Zettle (2003), Gaudiano and Herbert (2006), Woods et al. (2006), Gratz and Gunderson (2006), and Gregg, Callaghan, Hayes, and Glenn-Lawson (2007) only used one therapist. The remaining four ACT studies did not provide any information regarding the number of therapists.

In the DBT studies, however, only Safer, Telch, and Agras (2001) used a single therapist. Telch, Agras, and Linehan (2001) and Soler et al. (2005) used 2, Turner (2000) used 4, Linehan et al. (1991, 1999, 2002) and Koons et al. (2001) used 5, Lynch et al. (2003) used 6, whereas Linehan et al. (2006) and Verheul et al. (2003) used 16. Simpson et al. (2004) and Lynch et al. (2007) did not state the exact number of therapists, but it is evident that they used more than one.

Studies with just one therapist and no information studies were combined into one category and compared with multiple therapist studies. From this it was evident that ACT studies (15%) had significantly fewer (Fisher's exact probability tests, $P_2 = 0.0002$) multiple therapist studies than DBT (92%).

Specific problems pertain to the studies using TAU as comparison condition. First, the articles usually give no information (Dahl, Wilson, & Nilsson, 2004; Linehan et al., 1991, 1999), or very brief information (Bach & Hayes, 2002; Gregg et al., 2007; Verheul et al., 2003) on the therapists who carry out TAU. Somewhat more information is given in a few studies (Gaudiano & Herbert, 2006; Gratz & Gunderson, 2006; Koons et al., 2001) describing the number of therapists, their profession, and place of work. Only the study by Linehan et al. (2006) gives detailed information on the therapist providing TAU. A second problem is that in the studies that provided information on the amount of therapy patients in TAU received less therapy than those receiving DBT. Finally, in some studies TAU was carried out by the same therapists who referred the patients for DBT in the first place (e.g. Linehan et al., 1999; Verheul et al., 2003). This means that their motivation to do a good job with the patients, who apparently did not benefit from their treatments in the first place, can be questioned.

Adherence and competence

In order to be able to conclude that the treatment the authors describe they applied really was used, it is necessary to have independent assessors rating a randomized proportion (ideally 20%) of the recorded sessions for adherence to the treatment manual. Only two of the ACT studies (Hayes, Wilson, et al., 2004; Zettle & Rains, 1989) and four of the DBT studies (Koons et al., 2001; Linehan et al., 2006; Turner, 2000; Verheul et al., 2003) reported any form of adherence ratings. The only CBASP study and both IBCT studies had adherence ratings.

The picture looks even worse regarding competence ratings; no ACT, DBT, or CBASP study reported having an independent expert on the respective treatment rating the therapists' competence when carrying out the treatment. However, both IBCT studies had competence ratings.

Take the Zettle (2003) study as an example of a study lacking adherence and competence ratings. The study compared ACT and systematic desensitization (SD) for college students with mathematics anxiety, and the author was the only therapist in both treatments. Since Zettle is one of the first therapists working with ACT (Zettle & Hayes, 1986), it is possible, or even probable, that he has more experience doing ACT than SD, that he favors ACT over SD, and that he does it more in adherence with the manual and more competently than he does SD. There is no way of knowing since the study did not include or report these ratings.

Meta-analysis

Controlled effect sizes (ES) were calculated as Cohen's *d*, but since this suffers from a slight upward bias when based on small samples (Lipsey & Wilson, 2001) it was transformed to Hedges' *g*. The distributions of ESs in ACT and DBT studies, respectively, were checked for outliers, but all ESs were found to lie within the $M \pm 2$SD range. The total ES for the third wave studies was 0.56 ($z = 4.87$, $p < 0.0001$) and 95% CI (0.33, 0.79). The test of heterogeneity showed a significant *Q*-value (166.58, $p < 0.0001$), indicating differences among the various therapies and control conditions used in the studies.

ACT studies

The result of the overall meta-analysis using a random model showed a mean ES of 0.68 ($z = 5.11$, $p < 0.0001$) and 95% CI (0.42, 0.94). The test of heterogeneity yielded a significant *Q*-value (30.99, $p = 0.006$). This was followed by a moderator analysis on type of control condition. A mixed effects analysis led to a nonsignificant *Q*-value (1.48), with the ESs shown in Table 7. They were all significantly different from zero. Using Cohen's rule-of-thumb, the WLC comparison showed a large ES, whereas the TAU and active treatment comparisons yielded moderate ESs.

A fail-safe *N* using Orwin's (1983) test was calculated. In order to bring the ES down to a trivial value of 0.20, the number of missing studies with non-significant ES has to be 65.

DBT studies

The result of the overall meta-analysis showed a mean ES of 0.58 ($z = 5.81$, $p < 0.0001$) and 95% CI (0.38, 0.77). The test of heterogeneity yielded a nonsignificant *Q*-value (15.42). This was followed by an analysis of type of control condition. A fixed effect analysis resulted in the ESs shown in Table 7, which are all significantly different from zero. The WLC comparison yielded a large ES, but in comparison to TAU and active treatments the ES was moderate.

A fail-safe *N* using Orwin's (1983) test was calculated. In order to bring the ES down to a trivial value of 0.20, the number of missing studies with non-significant ES has to be 49.

Comparison of ACT vs. DBT

Statistical analyses of the ESs achieved by ACT and DBT for the respective comparison conditions yielded no significant differences, which is understandable in light of the small *n*'s.

Uncontrolled ES

The ES for within-group change on the primary measure was possible to calculate for 11 of the 13 studies in both ACT and DBT. At follow-up, however, this was only possible for eight ACT and four DBT studies. A check for outliers found one among the DBT studies, and this was reduced to $M + 2$SD before analysis. At post-treatment the mean ES was 1.04 for ACT (SD = 0.72) and 1.18 (SD = 0.89) for DBT. At follow-up the means were 1.13 (SD = 0.82) for ACT and 0.82 (SD = 0.87) for DBT (Table 7).

Table 7
Effect sizes for ACT, and DBT, studies divided on type of comparison condition

Comparison condition	ACT			DBT		
	N	ES	*Z*	*N*	ES	*Z*
Overall	15	0.68	5.11[c]	13	0.58	5.81[c]
Waiting list control	2	0.96	2.91[b]	2	1.30	4.67[c]
Treatment as usual	5	0.79	3.71[c]	4	0.47	2.40[a]
Active treatment	8	0.53	2.61[b]	7	0.47	3.74[c]

[a]$p<0.05$, [b]$p<0.01$, [c]$p<0.001$.

Are the third wave therapies empirically supported treatments?

This section evaluates the third wave therapies in relation to criteria I, and III–V of the Task Force (Chambless et al., 1998). The reason why criterion II is not included is that it concerns single-case design studies and these were excluded in the present review.

Criterion I:A

The first EST criterion reads: I. At least two *good* between-group design experiments must demonstrate efficacy in one or more of the following ways. Good in this instance means methodologically rigorous (Chambless et al., 1998; Chambless & Ollendick, 2001). This issue is the focus of the following sections.

ACT

Two of the ACT studies contained a comparison with a psychological placebo. Bond and Bunce (2000) compared ACT to the Innovation Promotion Program for stressed people in a media organization and found that ACT did better. However, this is a study of participants without a psychiatric disorder, the change on the primary measure is small, and there are numerous methodological weaknesses, e.g. no information on therapist training, adherence or competence ratings. Lundgren et al. (2006) compared ACT with supportive therapy (ST) for patients with epilepsy and there are some major methodological problems. ACT was combined with behavioral seizure control techniques, which in previous research (e.g. Dahl, Brorson, & Melin, 1992) have been shown to be effective, and there is no way to tease out what effect ACT has on the outcome. There is no indication that the patients experienced ST as a valid treatment for epilepsy, since no credibility ratings were done. All patients were institutionalized or day workers at a center for epilepsy in South Africa, which means that it is quite possible that the two groups of patients could talk to each other and compare the treatments they received. It is not far fetched to assume that the ST patients experienced what they got more as something inert than as an active treatment. Finally, there is no description of the therapists in the study; whether the same therapist treated both groups, what training and experience they had, etc.

Five studies compared ACT to another active treatment. Zettle and Hayes (1986) compared ACT with CT for depressed women, but the study is very briefly described, and has a number of methodological problems. The participants have not been diagnosed, the sample is not described, a manual has not been used, and on the primary measure the groups did not differ significantly at post-treatment, only at the 3-month follow-up. The study by Zettle and Rains (1989) is considerably better methodologically, but the depressed subjects have not been diagnosed, only one therapist (the first author) treated all patients in the three conditions, and there is no check of therapist competence. The third study by Zettle (2003) focused on college students with "mathemathics anxiety", which probably is not a psychiatric diagnosis (no diagnostic system was used). ACT was compared to systematic desensitization (SD), but the inclusion criteria are unclear, and only one therapist (the author) treated all participants. Furthermore, the groups did not differ on the primary measure, but on a secondary measure SD was significantly better than ACT. Gifford et al. (2004) compared ACT and NRT in smoking cessation. No diagnostic system was used, but the study has many strengths. However, the attrition rate was 36%, and the outcome of ACT was not better than NRT at post-treatment. At the 1-yr follow-up ACT did better than NRT in the completer sample, but not in the intent-to-treat sample. However, there is no information about any "uncontrolled" treatment that the subjects may have obtained during the follow-up year. This study did not show that ACT was better than the compared treatment. Hayes, Wilson, et al. (2004) studied a sample of opiate addicts who had ongoing MM treatment. In addition, one group got ACT and another got ITSF. This study has a number of strengths, but it suffers from a large attrition rate (37%). The three conditions did not differ significantly at post-treatment on the primary measure (percent clean urine analyses), but at the 6-month follow-up ACT was significantly better than MM both on opiates and on total drugs. However, ACT was still not better than ITSF at follow-up. Thus, the outcome mirrors that of Gifford et al., with no difference at post but a significant difference at follow-up. However, any possible treatment during the follow-up period was not assessed or not described, and ACT did not show a better outcome than the compared treatment.

DBT

Six studies compared DBT in some way to another active treatment. However, in two of these both groups got DBT and the independent variable was fluoxetine or placebo (Simpson et al., 2004) and olanzapine or placebo (Soler et al., 2005), respectively. The two studies by Lynch et al. (2003, 2007) compared the combination of DBT + ADM with ADM only, but obtained no significant differences. In none of these studies was DBT a single treatment that was compared to another treatment, and thus they cannot be used to ascertain whether DBT fulfils criterion I:A. Turner (2000) compared DBT to client centered therapy (CCT). The methodological problems are that DBT was combined with psychodynamic techniques to an unknown extent, and that the therapy varied between 49 and 84 sessions, with no information about the mean for each group. The sample was only briefly described and there were some unclear issues about the treatment programs. The last study is the recent one by Linehan et al. (2006) comparing DBT with CTBE. The authors go to a great length in order to describe the recruitment and characteristics of the CTBE therapists, which is very good. However, it is practically impossible to say what kind of therapy the patients in the CTBE condition received. The therapists described themselves as "eclectic but nonbehavioral" or "mostly psychodynamic" (p. 760), but it is not possible to say if this is one of the psychodynamic therapies that has been tested for BPD (Leichsenring & Leibing, 2003) or something else. There are no recordings of the sessions in CTBE, which makes it impossible to know if these therapists were adhering to a therapy protocol, or carrying out their treatment as competently as the DBT therapists did. Furthermore, as described above, the groups differ in the amount of therapy given, which in itself is a threat to internal validity. This study fulfils with some hesitation criterion I:A, since DBT was superior to CTBE on some of the central measures.

CBASP

The only CBASP study is a comparison between this treatment, ADM, and the combination of CBASP and ADM (Keller et al., 2000). It found no difference between CBASP and ADM, whereas the combination was significantly better than the individual treatments. This study has many good methodological qualities, but there was no information regarding control of concomitant treatments and follow-up data were not presented.

IBCT

The two IBCT studies also have a number of methodological strengths; however, there are also some weaknesses. Both lack independent blind evaluators and power analysis. Jacobson, Christensen, Prince, Cordova, and Eldridge (2000) also lack follow-up and control of concomitant treatments. Both studies found that there was no significant difference between IBCT and traditional behavioral couple therapy (TBCT), even if the there was a trend in favor of IBCT.

Criterion I:B

Criterion I:B stipulates *Equivalence* to already established treatment with adequate sample sizes ($n = 30$). This requires that the authors use a detailed equivalence analysis (e.g. Rogers, Howard, & Vessey, 1993; Schuirmann, 1987) in order to test if their treatment and the already established treatment yielded sufficiently similar results in order to be considered statistically equivalent. Among the ACT studies, Zettle and Hayes (1986), Zettle and Rains (1989), Zettle (2003), Gifford et al. (2004), and Hayes, Wilson, et al. (2004) did not find significant differences between ACT and the comparison treatment. However, none of these studies had large enough cell sizes to perform an equivalence analysis. Among DBT studies, Simpson et al. (2004), Soler et al. (2005), Lynch et al. (2003), and Lynch et al. (2007) also found no significant differences between DBT and the compared treatment. These studies also had cell sizes that were too small for an equivalence analysis. The only CBASP study (Keller et al., 2000) reported no significant difference between CBASP and ADM, but despite cell sizes well above 200 no equivalence analysis was done. Finally, both IBCT studies found non-significant differences compared to TBCT. Christensen et al. (2004) had large enough cell sizes to allow and equivalence analysis, but none was reported.

Criterion III and IV

Criterion III reads: Experiments must be conducted with treatment manuals or equivalent clear description of treatment. This criterion was fulfilled by all but one of the ACT studies (Zettle & Hayes, 1986).

Criterion IV stipulates thus: Characteristics of samples must be specified. This was fulfilled by all studies, except for two ACT studies (Bond & Bunce, 2000; Zettle & Hayes, 1986).

Summary of EST criteria and third wave studies

Table 8 gives a summary of my evaluation of the third wave therapy studies in relation to the different EST criteria. Regarding ACT, two studies (Bond & Bunce, 2000; Lundgren et al., 2006) found significantly better effects than a psychotherapy placebo condition. Four other studies (Bach & Hayes, 2002; Dahl et al., 2004; Gratz & Gunderson, 2006; Gregg et al., 2007) found ACT to be significantly better than TAU. However, all these studies had various methodological problems, as described above, which means that they cannot be considered good, i.e. methodologically rigorous, as the first criterion stipulates. Unfortunately, the methodologically best studies (e.g. Gifford et al, 2004; Hayes, Wilson, et al., 2004) did not find ACT to be superior to the comparison treatment and did not do an equivalence analysis.

When it comes to DBT, one study (Linehan et al., 1991) found better effects than for TAU, and another found better effects than for CCT (Turner, 2000). Both of these have methodological shortcomings, however, and are not good in the EST meaning. Finally, in the recent study of Linehan et al. (2006) both reported better effects than the comparison therapy and had a good methodology, thus fulfilling criteria I–IV. Since criterion V reads "Effects must be demonstrated by at least two investigators or teams", there needs to be a second study by another researcher (independent of Linehan) that fulfils all criteria before DBT can be considered as an EST. Regarding CBASP there is only one study, which does not fulfil criterion I.

Discussion

The purposes of this review article of third wave treatment RCTs were: (1) to describe and review them methodologically, (2) to meta-analytically assess their efficacy, and (3) to do a preliminary evaluation if they currently fulfil the criteria for ESTs.

The development of outcome research for a new form of therapy usually starts with a systematic case series, continues with small RCTs comparing the new treatment with WLC or TAU, and finally large-scale RCTs comparing with a previously shown effective treatment for the disorder in question. Thus, a young therapy usually does not have as much research, especially of the more advanced kind, as an older form of therapy, and this is reflected in the collection of third wave therapy studies. However, one may wonder how long a therapy can be said to be "young and promising". If we look at when the manuals for the various third wave

Table 8
Summary of the third wave studies in relation to the EST criteria

Study	Comparison condition	I:A. Superiority to pill or psychotherapy placebo	I:A. Superiority to other treatment	I:B. Equivalence to already established treatment	III. Treatment manuals	IV. Characteristics of samples must be specified
ACT studies						
Zettle (1986)	CT		=	0	–	–
Zettle (1989)	CT		=	0	+	+
Bond (2000)	IPP	>			+	–
Bach (2002)	TAU		>		+	+
Zettle (2003)	SD		=	0	+	+
Dahl (2004)	TAU		>		+	+
Gifford (2004)	NRT		=	0	+	+
Hayes (2004)	ITSF		=	0	+	+
Gaudiano (2006)	TAU		=		+	+
Woods (2006)	WLC	NA	NA		+	+
Gratz (2006)	TAU		>		+	+
Lundgren (2006)	ST	>			+	+
Gregg (2007)	TAU		>		+	+
DBT studies						
Linehan (1991)	TAU		>		+	+
Linehan (1999)	TAU		= (>)		+	+
Linehan (2002)	CVT		= (>)		+	+
Linehan (2006)	CTBE		>		+	+
Turner (2000)	CCT		>		+	+
Koons (2001)	TAU		=		+	+
Verheul (2003)	TAU		=		+	+
Simpson (2004)	Fluoxetine		=	0	+	+
Soler (2005)	Olanzepine		=	0	+	+
Safer (2001)	WLC	NA	NA		+	+
Telch (2001)	WLC	NA	NA		+	+
Lynch (2003)	ADM		=		+	+
Lynch (2007)	ADM		=		+	+
CBASP study						
Keller (2000)	ADM, Comb.		=	–	+	+
IBCT studies						
Jacobson (2000)	TBCT		=		+	+
Christensen (2004)	TBCT		=	–	+	+

CT = cognitive therapy, IPP = Innovation Promotion Program, TAU = treatment as usual, SD = systematic desensitization, NRT = nicotine replacement treatment, ITSF = Intensive Twelve Step Facilitation Program, WLC = waiting list control, ST = supportive therapy, CVT = comprehensive validation therapy, CTBE = community treatment by experts, CCT = client centred therapy, ADM = antidepressant medication, TBCT = traditional behavioral couple therapy. > = significantlty better than the comparison condition, = no significant difference between conditions, NA = not applicable, 0 = cell size too small to allow equivalence analysis, + = criterion is fulfilled, – = criterion is not fulfilled.

therapies were published we find the following: FAP 1991, DBT 1993, IBCT 1996, ACT 1999, and CBASP 2000. The mean number of RCTs per year since the publication of the respective manuals are: FAP 0, DBT 0.9, IBCT 0.2, ACT 1.6, and CBASP 0.2. It is difficult to compare the third wave and traditional CBT on this variable, but it is fair to conclude that this is not a particularly high publication rate.

Descriptive and methodological review

I am not aware of a review paper that compares two forms of therapy on methodological grounds and used a matched sample of studies published in the same journals and in the same year (±1 yr). However, matched control groups are often used in clinical research (Kazdin, 2003), where subjects are matched on e.g. gender and age or diagnosis and duration. The current review just used an adaptation of the matched control group design with studies instead of patients.

It is a well-known fact that journal editors often suggest page reductions before accepting an article. Sometimes this means that good methodological aspects of studies might be deleted and the printed study looks less stringent than it actually is. However, there is no reason to believe that this factor influences third wave and CBT studies differently, and thus, we can only evaluate the printed articles.

The descriptive review showed that only half of the ACT studies diagnosed their participants, whereas this was done in all DBT, CBASP, and CBT studies. This is difficult to understand, since there does not seem to be an ideological resistance to diagnosing among ACT researchers. Further, the number of participants starting DBT studies was lower and the attrition rate higher than in CBT studies. Since most DBT studies focus on BPD and the prevalence of this disorder is 2% in the general population and 10% in outpatient mental health clinics (APA, 2000), there should not be a shortage of suitable patients. The DBT studies also had a higher proportion of females than CBT studies, which is understandable given that most studies were done on BPD that has a preponderance of females.

The analysis of methodological stringency showed that the third wave studies had significantly lower mean scores than CBT studies, and that both ACT and DBT studies had lower means than CBT studies, when year of publication and journal were controlled for. Furthermore, this difference was not caused by the 11 CBT studies in non-matching journals having that much higher score than their third wave counterparts. The 11 items on which either one or both third wave treatments' studies had lower scores than CBT studies were the following: 3—representativeness of the sample, 4—reliability of the diagnosis, 6—reliability and validity of outcome measures, 9—assignment to treatments, 14—number of therapists, 15—therapist training/experience, 16—checks for treatment adherence, 18—control of concomitant treatments, 20—statistical analyses and presentation of data, 21—clinical significance, and 22—equality of therapy hours. It is difficult to understand why the third wave studies had lower scores on some of these items, e.g. reliability and validity of outcome measures, number of therapists and their training/experience, statistical analyses and presentation of results, and equality of therapy hours. It does not require a higher budget to use the best outcome measures than having untested measures; to divide the patients to two or more therapists, instead of just having one; to analyze and present the results in a complete way; or to make sure that the compared active treatments get equal hours of therapy. Other items, on the other hand, take a toll in the budget, e.g. reliability testing of the diagnosis and checks for treatment adherence, and one can understand if a low budget study chooses not to do it. However, there was no difference between third wave and CBT studies regarding the frequency of projects receiving grants, but it was not possible to decide if the size of the grants was large enough to satisfy these methodological features.

Meta-analysis

The meta-analysis yielded mean ESs in the moderate range for both ACT (0.68) and DBT (0.58). The ES for ACT is very similar to that (0.66) obtained by Hayes, Luoma, Bond, Masuda, and Lillis (2006), which included all studies in the present meta-analysis together with four unpublished or non-disorder studies which were excluded in the present review. When the active treatments were compared to waitlist conditions the ESs were large, 0.96 and 1.30, respectively, but in both cases only two studies made up the ES. The comparisons with TAU and another active treatment, respectively, resulted in moderate ESs for both therapies. All ESs were significantly different from zero and the fail-safe analysis indicated that it would take 65 unpublished ACT studies and 49 unpublished DBT studies to bring the ESs down to an insignificant level. It is hard to imagine that there are 3–4 times as many unpublished as already published third wave studies hidden in file drawers. Thus, it is safe to conclude that the third wave studies yield a significant, but moderate effect size.

Criteria for EST

It must, of course, be acknowledged that the EST criteria are not unproblematic, but have given rise to a lot of debate since they were published (e.g. Herbert, 2003). Besides suggesting to delete the criteria altogether because it is a bad idea, I do not believe that there has been published an alternative set of criteria that are clearly better than the original Task Force criteria of 1995.

The review of the RCTs for ACT shows that there is no study that is methodologically good enough and also fulfils the first of the EST criteria. When it comes to DBT there is one study (Linehan et al., 2006) that can be considered fulfilling the first criterion. This study also fulfils the third criterion (using a treatment manual) and the fourth (characteristics of the sample are specified). However, in order for DBT to be considered a well-established treatment there has to be a second, methodologically good RCT, by a different investigator/team, showing that DBT is superior to a psychotherapy placebo, or to another treatment, or equivalent to an already established treatment. The DBT studies on BPD published by researchers other than Linehan have different methodological problems (e.g. an astonishingly high dropout rate of 59% in Verheul et al., 2003; a small and restricted sample in Koons et al., 2001; a combination of DBT and psychodynamic therapy in Turner, 2000), and thus criterion I and V are not fulfilled.

None of the CBASP and IBCT studies fulfil the EST criteria. However, both the CBASP study of Keller et al. (2000) and the IBCT study of Christensen et al. (2004) would do so had they performed an equivalence analysis, providing it showed that the treatments in these studies were statistically equivalent.

Limitations

A review and meta-analysis is of course limited by the number and quality of the studies that are included. In the present review 29 RCTs on five of the third wave treatments could be included, which is enough for a meta-analysis (at least regarding the overall ES).

The 10 variables used for the descriptive analysis yielded a grid of 130 cells for both ACT and DBT. There were 8 no information cells for ACT (6.2%) and 8 for DBT (6.2%), and 1 (3%) for the other therapies combined, which means that the results are valid in this respect.

To the best of my knowledge there is no established and well-known instrument to evaluate the methodological quality of psychotherapy outcome studies. The scale used for assessing the methodological stringency was developed for this article, and it had good internal consistency and inter-rater reliability. One could always discuss the content of a rating scale, and have various opinions of which items should be included or not. The scale used in this article was based on a previous one by Tolin (1999), which was an operationalization of the Foa and Meadows (1997) "golden standard" for therapy outcome studies in PTSD. If certain items that may be important have been left out unintentionally, there is no reason to believe that these are items where the third wave studies are particularly strong and CBT studies particularly weak, so that an inclusion of them would result in equal mean scores for third wave and CBT studies. Since the mean for third wave studies was 19.6 and that for CBT studies 27.8, it would take four new methodology items on which all third wave studies scored a 2 and all CBT studies a 0, in order for the means to be equal, and three items to no longer differ significantly. It is equally probable that CBT studies would be strong in those items as well and the significant difference would remain. Thus, I believe that the rating scale used in this article gives a fair description of the methodological stringency in third wave studies so far.

Recommendation for future research

If the proponents of the third wave treatments want to see their therapies on the list of ESTs, which is a fair assumption, there are some suggestions they should consider when planning future RCTs. The list below should not be perceived to indicate that the third wave studies all lack in these respects; some do and some do not, but by incorporating these suggestions the chance of being listed as an EST in the future will increase, providing that the outcome is good of course:

(1) Do not use WLC as the control condition, since criterion I requires a placebo or another treatment.
(2) Do not use TAU as the control condition, since the methodological problems described above are so extensive.
(3) Use an active treatment as comparison, preferably one that has been established as effective for the disorder in question.
(4) Do a proper power analysis before the start of the study and adjust the cell size for the attrition that may occur.
(5) Use a representative sample of patients, diagnose them using suitable instruments in the hands of trained interviewers, and test the diagnostic reliability.
(6) Let an independent researcher or agency use an unobjectionable randomization procedure, and conceal the outcome of it from all persons involved in the study.
(7) Use reliable and valid outcome measures; both the ones that are specific to the disorder and general ones.
(8) Use blind assessors and evaluate their blindness regarding treatment condition of the patients they assess.
(9) Train the assessors properly and measure inter-rater reliability on the data collected throughout the study (not just during training).
(10) Use three or more properly trained therapists and randomize patients to therapist to enable an analysis of possible therapist effect on the outcome.
(11) Include at least a 1-year follow-up in the study and assess any nonprotocol treatments that the patients may have obtained during the follow-up period.
(12) Audio- or videotape all therapy sessions. Randomly select 20% of these and let independent experts rate adherence to treatment manual and therapist competence.
(13) Insert procedures to control for concomitant treatments that patients in the study may obtain simultaneously as the protocol treatment.
(14) Describe the attrition, do a drop-out analysis and include all randomized subjects in an intent-to-treat analysis.
(15) Assess clinical significance of the improvement of the primary measures.

Regarding the focus of future research, my recommendations are more therapy specific:

- ACT should be compared with CBT for the most common psychiatric disorders, e.g. the various anxiety disorders, depression, and eating disorders.
- DBT should be compared with psychodynamic therapy (the version with the strongest evidence base) for borderline PD, and male subjects ought to be included. When it comes to eating disorders DBT should be compared to Fairburn's CBT (Fairburn, Cooper, & Shafran, 2003), and for depression with CBT or behavioral activation (Martell, Addis, & Jacobson, 2001).
- CBASP should be compared with CBT (e.g. DeRubeis, Hollon, Amsterdam, Shelton, & Young, 2005) for chronic and other forms of depression.
- FAP should be subjected to RCTs for depression and then for other common disorders.
- IBCT should be evaluated in good RCTs by researchers who are independent from the originators of this therapy.

Conclusions

The following conclusions can be drawn from the present review and meta-analysis. First, the third wave treatment RCTs published so far have used a research methodology that is significantly less stringent than CBT studies published during the same years and in the same journals. Second, the mean effect size was moderate for both ACT and DBT studies. Third, at this time no third wave therapy fulfils the criteria for empirically supported treatments.

Acknowledgements

Portions of this article were presented at the Second World Conference on ACT, RFT, and Contextual Behavioral Science, London, July 24–28, 2006. The help of Lotta Reuterskiöld for assisting with the inter-rater reliability testing, Ata Ghaderi, Lennart Melin and Jonas Ramnerö for valuable comments on the manuscript, and Steven Hayes for continuously sending me manuscripts of submitted and accepted articles on ACT is gratefully acknowledged.

Appendix A. Psychotherapy outcome study methodology rating form[3]

Note: If not enough information is given regarding a specific item a rating of 0 is given.

1. Clarity of sample description

   0 *Poor*. Vague description of sample (e.g. only mentioned whether patients were diagnosed with the disorder).

   1 *Fair*. Fair description of sample (e.g. mentioned inclusion/exclusion criteria, demographics, etc.).

   2 *Good*. Good description of sample (e.g. mentioned inclusion/exclusion criteria, demographics, and the prevalence of comorbid disorders).

2. Severity/chronicity of the disorder

   0 *Poor*. Severity/chronicity was not reported and/or subsyndromal patients were included in the sample.

   1 *Fair*. All patients met the criteria for the disorder. Sample includes acute (<1 yr) and/or low severity.

   2 *Good*. Sample consisted entirely of chronic (>1 yr) patients of at least moderate severity.

3. Representativeness of the sample

   0 *Poor*. Sample is very different from patients seeking treatment for the disorder (e.g. there are excessively strict exclusion criteria).

   1 *Fair*. Sample is somewhat representative of patients seeking treatment for the disorder (e.g. patients were only excluded if they met criteria for other major disorders).

   2 *Good*. Sample is very representative of patients seeking treatment for the disorder (e.g. authors made efforts to ensure representativeness of sample).

4. Reliability of the diagnosis in question

   0 *Poor*. The diagnostic process was not reported, or not assessed with structured interviews by a trained interviewer.

   1 *Fair*. The diagnosis was assessed with structured interview by a trained interviewer.

   2 *Good*. The diagnosis was assessed with structured interview by a trained interviewer and adequate inter-rater reliability was demonstrated (e.g. *kappa* coefficient).

5. Specificity of outcome measures

   0 *Poor*. Very broad outcome measures, not specific to the disorder (e.g. SCL-90R total score).

   1 *Fair*. Moderately specific outcome measures.

   2 *Good*. Specific outcome measures, such as a measure for each symptom cluster.

6. Reliability and validity of outcome measures

   0 *Poor*. Measures have unknown psychometric properties, or properties that fail to meet current standards of acceptability.

   1 *Fair*. Some, but not all measures have known or adequate psychometric properties.

   2 *Good*. All measures have good psychometric properties. The outcome measures are the best available for the authors' purpose.

[3] Modified after and substantially elaborated (10 new items): Tolin (1999).

7. Use of blind evaluators

   0 *Poor*. Blind assessor was not used (e.g. assessor was the therapist, assessor was not blind to treatment condition, or the authors do not specify).

   1 *Fair*. Blind assessor was used, but no checks were used to assess the blind.

   2 *Good*. Blind assessor was used in correct fashion. Checks were used to assess whether the assessor was aware of treatment condition.

8. Assessor training

   0 *Poor*. Assessor training and accuracy are not specified, or are unacceptable.

   1 *Fair*. Minimum criterion for assessor training is specified (e.g. assessor has had specific training in the use of the outcome measure), but accuracy is not monitored or reported.

   2 *Good*. Minimum criterion of assessor training is specified. Inter-rater reliability was checked, and/or assessment procedures were calibrated during the study to prevent evaluator drift.

9. Assignment to treatment

   0 *Poor*. Biased assignment, e.g. patients selected their own therapy or were assigned in another non-random fashion, or there is only one group.

   1 *Fair*. Random or stratified assignment. There may be some systematic bias but not enough to pose a serious threat to internal validity. There may be therapist by treatment confounds. *N* may be too small to protect against bias.

   2 *Good*. Random or stratified assignment, and patients are randomly assigned to therapists within condition. When theoretically different treatments are used, each treatment is provided by a large enough number of different therapists. *N* is large enough to protect against bias.

10. Design

    0 *Poor*. Active treatment vs. WLC, or briefly described TAU.

    1 *Fair*. Active treatment vs. TAU with good description, or placebo condition.

    2 *Good*. Active treatment vs. another previously empirically documented active treatment.

11. Power analysis

    0 *Poor*. No power analysis was made prior to the initiation of the study.

    1 *Fair*. A power analysis based on an estimated effect size was used.

    2 *Good*. A data-informed power analysis was made and the sample size was decided accordingly.

12. Assessment points

    0 *Poor*. Only pre- and post-treatment, or pre- and follow-up.

    1 *Fair*. Pre-, post-, and follow-up <1 year.

    2 *Good*. Pre-, post-, and follow-up $\geqslant 1$ year.

13. Manualized, replicable, specific treatment programs

    0 *Poor*. Description of treatment procedure is unclear, and treatment is not based on a publicly available, detailed treatment manual. Patients may be receiving multiple forms of treatment at once in an uncontrolled manner.

    1 *Fair*. Treatment is not designed for the disorder, or description of the treatment is generally clear and based on a publicly available, detailed treatment manual, but there are some ambiguities about the procedure. Patients may have received additional forms of treatment, but this is balanced between groups or otherwise controlled.

    2 *Good*. Treatment is designed for the disorder. A detailed treatment manual is available, and/or treatment is explained in sufficient detail for replication. No ambiguities about the treatment procedure. Patients receive only the treatment in question.

14. Number of therapists

    0 *Poor*. Only one therapist, i.e. complete confounding between therapy and therapist.

    1 *Fair*. At least two therapists, but the effect of therapist on outcome is not analyzed.

    2 *Good*. Three, or more therapists, and the effect of therapist on outcome is analyzed.

15. Therapist training/experience

    0 *Poor*. Very limited clinical experience of the treatment and/or disorder (e.g. students).

    1 *Fair*. Some clinical experience of the treatment and/or disorder.

    2 *Good*. Long clinical experience of the treatment and the disorder (e.g. practicing therapists).

16. Checks for treatment adherence

    0 *Poor*. No checks were made to assure that the intervention was consistent with protocol.

    1 *Fair*. Some checks were made (e.g. assessed a proportion of therapy tapes).

    2 *Good*. Frequent checks were made (e.g. weekly supervision of each session using a detailed rating form).

17. Checks for therapist competence

    0 *Poor*. No checks were made to assure that the intervention was delivered competently.

    1 *Fair*. Some checks were made (e.g. assessed a proportion of therapy tapes).

    2 *Good*. Frequent checks were made (e.g. weekly supervision of each session using a detailed rating form).

18. Control of concomitant treatments (e.g. medications)

    0 *Poor*. No attempt to control for concomitant treatments, or no information about concomitant treatments provided. Patients may have been receiving other forms of treatment in addition to the study treatment.

    1 *Fair*. Asked patients to keep medications stable and/or to discontinue other psychological therapies during the treatment.

    2 *Good*. Ensured that patients did not receive any other treatments (medical or psychological) during the study.

19. Handling of attrition

    0 *Poor*. Proportions of attrition are not described, or described but no dropout analysis is performed.

    1 *Fair*. Proportions of attrition are described, and dropout analysis or intent-to-treat analysis is performed.

    2 *Good*. No attrition, or proportions of attrition are described, dropout analysis is performed, and results are presented as intent-to-treat analysis.

20. Statistical analyses and presentation of results

    0 *Poor*. Inadequate statistical methods are used and/or data are not fully presented.

    1 *Fair*. Adequate statistical methods are used but data are not fully presented.

    2 *Good*. Adequate statistical methods are used and data are presented with *M* and SD.

21. Clinical significance

    0 *Poor*. No presentation of clinical significance was done.

    1 *Fair*. An arbitrary criterion for clinical significance was used and the conditions were compared regarding percent clinically improved.

    2 *Good*. Jacobson's criteria for clinical significance were used and presented for a selection (or all) of the outcome measures, and conditions were compared regarding percent clinically improved.

22. Equality of therapy hours (for non-WLC designs only)

    0 *Poor*. Conditions differ markedly (⩾20% difference in therapy hours).

    1 *Fair*. Conditions differ somewhat (10–19% difference in therapy hours).

    2 *Good*. Conditions do not differ (<10% difference in therapy hours).

Appendix B. CBT studies used in the comparison

Alexopoulos, G. S., Raue, P., & Areán, P. (2003). Problem-solving therapy versus supportive therapy for geriatric major depression with executive dysfunction. *American Journal of Geriatic Psychiatry*, *11*, 46–52.

Arntz, A. (2002). Cognitive therapy versus interoceptive exposure as treatment of panic disorder without agoraphobia. *Behaviour Research and Therapy*, *40*, 325–341.

Barlow, D. H., Craske, M. G., Cerny, J. A., & Klosko, J. S. (1989). Behavioral treatment of panic disorder. *Behavior Therapy*, *20*, 261–282.

Barrowclough, C., Haddock, G., Tarrier, N., Lewis, S. W., Moring, J., O'Brien, R. et al. (2001). Randomized controlled trial of motivational interviewing, cognitive behavior therapy, and family intervention for patients with comorbid schizophrenia and substance use disorders. *American Journal of Psychiatry*, *158*, 1706–1713.

Bastien, C. H., Morin, C. M., Ouellet. M.-C., Blais, F. C., & Bouchard, S. (2004). Cognitive-behavioral therapy for insomnia: Comparison of individual therapy, group therapy, and telephone consultations. *Journal of Consulting and Clinical Psychology*, *72*, 653–659.

Bellack, A. S., Bennett, M. E., Gearon, J. S., Brown, C. H., & Yang, Y. (2006). A randomized clinical trial of a new behavioral treatment for drug abuse in people with severe and persistent mental illness. *Archives of General Psychiatry*, *63*, 426–432.

Blanchard, E. B., Hickling, E. J., Devineni, T., Veasey, C. H., Galovski, T. E., Mundy, E. et al. (2003). A controlled evaluation of cognitive behavioral therapy for posttraumatic stress in motor vehicle accident survivors. *Behaviour Research and Therapy*, *41*, 79–96.

Boelen, P. A., de Keijser, J., van den Hout, M. A., & van den Bout, J. (2007). Treatment of complicated grief: A comparison between cognitive-behavioral therapy and supportive counselling. *Journal of Consulting and Clinical Psychology*, 75, 277–284.

Bögels, S. M. (2006). Task concentration training versus applied relaxation, in combination with cognitive therapy, for social phobia patients with fear of blushing, trembling, and sweating. *Behaviour Research and Therapy*, *44*, 1199–1210.

Borkovec, T. D., Newman, M. G., Pincus, A. L., & Lytle, R. (2002). A component analysis of cognitive-behavioral therapy for generalized anxiety disorder and the role of interpersonal problems. *Journal of Consulting and Clinical Psychology*, *70*, 288–298.

Bryant, R. A., Moulds, M. L., Nixon, R. D. V., Mastrodomenico, J., Felmingham, K., & Hopwood, S. (2006). Hypnotherapy and cognitive behaviour therapy of acute stress disorder: A 3-year follow-up. *Behaviour Research and Therapy*, *44*, 1331–1335.

Cinciripini, P. M., Cinciripini, L. G., Wallfisch, A., Haque, W., & Vunakis, H. V. (1996). Behavior therapy and the transdermal nicotine patch: Effects on cessation outcome, affect, and coping. *Journal of Consulting and Clinical Psychology*, *64*, 314–323.

Clark, D. M., Ehlers, A., Hackman, A., McManus, F., Fennell, M., Grey, N. et al. (2006). Cognitive therapy versus exposure and applied relaxation in social phobia: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *74*, 568–578.

Craske, M. G., Lang, A. J., Aikins, D, & Mystkowski, J. L. (2005). Cognitive behavioral therapy for nocturnal panic. *Behavior Therapy*, *36*, 43–54.

Fairburn, C. G., Jones, R., Peveler, R. C., Carr, S. J., Solomon, R. A., O'Connor, M. E. et al. (1991). Three psychological treatments for bulimia nervosa. A comparative trial. *Archives of General Psychiatry*, *48*, 463–469.

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

Gilroy, L. J., Kirkby, K. C., Daniels, B. A., Menzies, R. G., & Montgomery, I. M. (2000). Controlled comparison of computer-aided vicarious exposure versus live exposure in the treatment of spider phobia. *Behavior Therapy*, *31*, 733–744.

Gruber, K., Moran, P. J., Roth, W. T., & Taylor, C. B. (2001). Computer-assisted cognitive behavioral group therapy for social phobia. *Behavior Therapy*, *32*, 155–165.

Herbert, J. D., Gaudiano, B. A., Rheingold, A. A., Myers, V. H., Dalrymple, K., & Nolan, E. M. (2005). Social skills training augments the effectiveness of cognitive behavioral group therapy for social anxiety disorder. *Behavior Therapy*, *36*, 125–138.

Kubany, E. S., Hill, E. E., Owens, J. A., Iannce-Spencer, C., McCaig, M. A., Tremayne, K. J. et al. (2004). Cognitive trauma therapy for battered women with PTSD (CTT-BW). *Journal of Consulting and Clinical Psychology*, *72*, 3–18.

Ladouceur, R., Dugas, M. J., Freeston, M. H., Léger, E., Gagnon, F., Thibodeau, N. (2000). Efficacy of a cognitive–behavioral treatment for generalized anxiety disorder: Evaluation in a controlled clinical trial. *Journal of Consulting and Clinical Psychology*, *68*, 957–968.

Loeb, K. L., Wilson, G. T., Gilbert, J. S., & Labouvie, E. (2000). Guided and unguided self-help for binge eating. *Behaviour Research and Therapy*, *38*, 259–272.

McLean, P. D., Whittal, M. L., Thordarson, D. S., Taylor, S., Söchting, I., Koch, W. J., et al. (2001). Cognitive versus behavior therapy in the group treatment of obsessive–compulsive disorder. *Journal of Consulting and Clinical Psychology*, *69*, 205–214.

Nauta, H., Hospers, H., Kok, G., & Jansen, A. (2000). A comparison between cognitive and behavioral treatment for obese binge eaters and obese non-binge eaters. *Behavior Therapy*, *31*, 441–461.

Rohan, K. J., Roecklein, K. A., Lindsey, K. T., Johnson, L. G., Lippy, R. D., Lacy, T. J., et al. (2007). A randomized controlled trial of cognitive-behavioral therapy, light therapy, and their combination for seasonal affective disorder. *Journal of Consulting and Clinical Psychology*, *75*, 489–500.

Rothbaum, B. O., Anderson, P., Zimand, E., Hodges, L., Lang, D., & Wilson, J. (2006). Virtual reality exposure therapy and standard (in vivo) exposure therapy in the treatment of fear of flying. *Behavior Therapy*, *37*, 80–90.

Tarrier, N., & Sommerfield, C. (2004). Treatment of chronic PTSD by cognitive therapy and exposure: 5-year follow-up. *Behavior Therapy*, *35*, 231–246.

Teri, L., & Lewinsohn, P. M. (1986). Individual and group treatment of unipolar depression: Comparison of treatment outcome and identification of predictors of successful treatment outcome. *Behavior Therapy*, *17*, 215–228.

Verduyn, C., Barrowclough, C., Roberts, J., Tarrier, N., & Harrington, R. (2003). Maternal depression and child behaviour problems. *British Journal of Psychiatry*, *183*, 342–348.

Wright, J. H., Wright, A. S., Albano, A. M., Basco, M. R., Goldsmith, L. J., Raffield, T., et al. (2005). Computer-assisted cognitive therapy for depression: Maintaining efficacy while reducing therapist time. *American Journal of Psychiatry*, *162*, 1158–1164.

References

American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed). Washington, DC: Author.

American Psychiatric Association (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev). Washington, DC: Author.

* Bach, P., & Hayes, S. C. (2002). The use of acceptance and commitment therapy to prevent the rehospitalization of psychotic patients: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *70*, 1129–1139.

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.

Bishop, S. R. (2002). What do we really know about mindfulness-based stress reduction? *Psychosomatic Medicine*, *64*, 71–84.

Biostat, Inc (2006). *Comprehensive Meta Analysis, version 2*. Englewood, NJ: Author.

* Bond, F. W., & Bunce, D. (2000). Mediators of change in emotion-focused and problem-focused worksite stress management interventions. *Journal of Occupational Health Psychology*, *5*, 156–163.

Borkovec, T. D., & Nau, S. D. (1972). Credibility of analogue therapy rationales. *Journal of Behavior Therapy and Experimental Psychiatry*, *3*, 257–260.

Chambless, D. L., Baker, M., Baucom, D. H., Beutler, L. E., Calhoun, K. S., et al. (1998). Update on empirically validated therapies, II. *The Clinical Psychologist*, *51*, 3–16.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716.

Corrigan, P. W. (2001). Getting ahead of the data: A threat to some behavior therapies. *The Behavior Therapist*, *24*, 189–193.

* Christensen, A., Atkins, D. C., Berns, S., Wheeler, J., Baucom, D. H., & Simpson, L. E. (2004). Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. *Journal of Consulting and Clinical Psychology*, *72*, 176–191.

Dahl, J., Brorson, L. O., & Melin, L. (1992). Effects of a broad-spectrum behavioral medicine treatment program on children with refractory epileptic seizures: An 8 year follow-up. *Epilepsia*, *33*, 98–102.

* Dahl, J., Wilson, K. G., & Nilsson, A. (2004). Acceptance and commitment therapy and the treatment of persons at risk for long-term disability resulting from stress and pain symptoms: A preliminary randomized trial. *Behavior Therapy*, *35*, 785–801.

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., et al. (2005). Cognitive therapy vs medications in the treatment of moderate to severe depression. *Archives of General Psychiatry*, *62*, 409–416.

Eysenck, H. (1952). The effects of psychotherapy. *Journal of Consulting Psychology*, *16*, 319–324.

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

Fairburn, C. G., Cooper, Z., & Shafran, R. (2003). Cognitive behaviour therapy for eating disorders: A "transdiagnostic" theory and treatment. *Behaviour Research and Therapy*, *41*, 509–528.

Feske, U., & Chambless, D. L. (1995). Cognitive behavioral versus exposure only treatment for social phobia: A meta-analysis. *Behavior Therapy*, *26*, 695–720.

Foa, E. B., & Meadows, E. A. (1997). Psychosocial treatments for posttraumatic stress disorder: A critical review. *Annual Review of Psychology*, *48*, 449–480.

* Gaudiano, B. A., & Herbert, J. D. (2006). Acute treatment of inpatients with psychotic symptoms using acceptance and commitment therapy: Pilot results. *Behaviour Research and Therapy*, *44*, 415–437.

Giesen-Bloo, J., van Dyck, R., Spinhoven, P., van Tilburg, W., Dirksen, C., et al. (2006). Outpatient psychotherapy for borderline personality disorder. Randomized trial of schema-focused therapy vs transference-focused psychotherapy. *Archives of General Psychiatry*, *63*, 649–658.

* Gifford, E. V., Kohlenberg, B. S., Hayes, S. C., Antonuccio, D. O., Piasecki, M. M., et al. (2004). Acceptance-based treatment for smoking cessation. *Behavior Therapy*, *35*, 689–705.

* Gratz, K. L., & Gunderson, J. G. (2006). Preliminary data on an acceptance-based emotion regulation group intervention for deliberate self-harm among women with borderline personality disorder. *Behavior Therapy*, *37*, 25–35.

* Gregg, J. A., Callaghan, G. M., Hayes, S. C., & Glenn-Lawson, J. L. (2007). Improving diabetes self-management through acceptance, mindfulness, and values: A randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *75*, 336–343.

Grossman, P., Niemann, L., Schmidt, S., & Walach, H. (2004). Mindfulness-based stress reduction and health benefits. A meta-analysis. *Journal of Psychosomatic Research*, *57*, 35–43.

Hayes, S. C. (2002). On being visited by the vita police: A reply to Corrigan. *The Behavior Therapist*, *25*, 134–137.

Hayes, S. C. (2004). Acceptance and commitment therapy, relational frame theory, and the third wave of behavioral and cognitive therapies. *Behavior Therapy*, *35*, 639–665.

Hayes, S. C., Luoma, J. B., Bond, F. W., Masuda, A., & Lillis, J. (2006). Acceptance and commitment therapy: Model, processes and outcomes. *Behaviour Research and Therapy*, *44*, 1–25.

Hayes, S. C., Masuda, A., Bissett, R., Luoma, J., & Guerrero, L. F. (2004). DBT, FAP, and ACT: How empirically oriented are the new behavior therapy technologies? *Behavior Therapy*, *35*, 3–54.

Hayes, S. C., Strosahl, K., & Wilson, K. G. (1999). *Acceptance and commitment therapy*. New York: Guilford Press.

* Hayes, S. C., Wilson, K. G., Gifford, E. V., Bissett, R., Piasecki, M., et al. (2004). A preliminary trial of twelve-step facilitation and acceptance and commitment therapy with polysubstance-abusing methadone-maintained opiate addicts. *Behavior Therapy*, *35*, 667–688.

Herbert, J. D. (2003). The science and practice of empirically supported treatments. *Behavior Modification*, *27*, 412–430.

Hofmann, S. G., & Asmundson, G. J. G. (2008). Acceptance and mindfulness-based therapy: New wave or old hat? *Clinical Psychology Review*, *28*, 1–16.

Jacobson, N. S., & Christensen, A. (1996). *Acceptance and change in couple therapy: A therapist's guide to transforming relationships*. New York: Norton.

* Jacobson, N. S., Christensen, A., Prince, S. E., Cordova, J., & Eldridge, K. (2000). Integrative behavioral couple therapy: An acceptance-based, promising new treatment for couple discord. *Journal of Consulting and Clinical Psychology*, *68*, 351–355.

Kabat-Zinn, J. (1990). *Full catastrophe living using the wisdom of your body and mind to face stress, pain and illness*. New York: Delacorte.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed). Boston: Allyn & Bacon.

* Keller, M. B., McCullough, J. P., Klein, D. N., Arnow, B., Dunner, D. L., et al. (2000). A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *The New England Journal of Medicine*, *342*, 1462–1470.

Kohlenberg, R. J., Kanter, J. W., Bolling, M. Y., Parker, C. R., & Tsai, M. (2002). Enhancing cognitive therapy for depression with functional analytic psychotherapy: Treatment guidelines and empirical findings. *Cognitive and Behavioral Practice*, *9*, 213–229.

Kohlenberg, R. J., & Tsai, M. (1991). *Functional analytic psychotherapy: Creating intense and curative therapeutic relationships*. New York: Plenum.

* Koons, C. R., Robins, C. J., Tweed, J. L., Lynch, T. R., Gonzalez, A. M., et al. (2001). Efficacy of dialectical behavior therapy in women veterans with borderline personality disorder. *Behavior Therapy*, *32*, 371–390.

Leichsenring, F., & Leibing, E. (2003). The effectiveness of psychodynamic therapy and cognitive behavior therapy in the treatment of personality disorders: A meta-analysis. *American Journal of Psychiatry*, *160*, 1223–1232.

Linehan, M. M. (1993). *Cognitive-behavioral treatment of borderline personality disorder*. New York: Guilford Press.

* Linehan, M. M., Armstrong, H. E., Suarez, A., Allmon, D., & Heard, H. (1991). Cognitive-behavioral treatment of chronically parasuicidal borderline patients. *Archives of General Psychiatry*, *48*, 1060–1064.

* Linehan, M. M., Comtois, K. A., Murray, A. M., Brown, M. Z., & Gallop, R. J. etal. (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, S. A., & Welch, S. S. etal. (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., Schmidt, H., Dimeff, L. A., Craft, C., Kanter, J., & Comtois, K. A. (1999). Dialectical behavior therapy for patients with borderline personality disorder and drug-dependence. *The American Journal on Addictions*, *8*, 279–292.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: SAGE Publications.

* Lundgren, T., Dahl, J., Melin, L., & Kies, B. (2006). Evaluation of acceptance and commitment therapy for drug refractory epilepsy: A randomized controlled trial in South Africa—a pilot study. *Epilepsia*, *47*, 2173–2179.

* Lynch, T. R., Morse, J. Q., Mendelson, T., & Robins, C. J. (2003). Dialectical behavior therapy for depressed older adults. *American Journal of Geriatric Psychiatry*, *11*, 33–45.

* Lynch, T. R., Cheavens, J. S., Cukrowicz, K. C., Thorp, S. R., Bronner, L., & Beyer, J. (2007). Treatment of older adults with co-morbid personality disorder and depression: A dialectical behavior therapy approach. *International Journal of Geriatric Psychiatry*, *22*, 131–143.

Martell, C. R., Addis, M. E., & Jacobson, N. S. (2001). *Depression in context: Strategies for guided action*. New York: Norton.

McCullough, J. P., Jr. (2000). *Treatment for chronic depression: Cognitive Behavioral Analysis System of Psychotherapy (CBASP)*. New York: Guilford Press.

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.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*, 157–159.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*, 553–565.

Roth, A., & Fonagy, P. (2005). *What works for whom? A critical review of psychotherapy research* (2nd ed). New York: Guilford Press.

* Safer, D. L., Telch, C. F., & Agras, W. S. (2001). Dialectical behavior therapy for bulimia nervosa. *American Journal of Psychiatry*, *158*, 632–634.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of bioavailability. *Journal of Pharmacokinetics and Biopharmaceptics*, *15*, 657–681.

Segal, Z. V., Williams, J. M. G., & Teasdale, J. T. (2001). *Mindfulness-based cognitive therapy for depression: A new approach to preventing relapse*. New York: Guilford Press.

* Simpson, E. B., Yen, S., Costello, E., Rosen, K., Begin, A., et al. (2004). Combined dialectical behavior therapy and fluoxetine in the treatment of borderline personality disorder. *Journal of Clinical Psychiatry*, *65*, 379–385.

* Soler, J., Pasqual, J. C., Campins, J., Barrachina, J., Puigdemont, D., et al. (2005). Double-blind, placebo-controlled study of dialectical behavior therapy plus olanzapine for borderline personality disorder. *American Journal of Psychiatry*, *162*, 1221–1224.

* Telch, C. F., Agras, W. S., & Linehan, M. M. (2001). Dialectical behavior therapy for binge eating disorder. *Journal of Consulting and Clinical Psychology*, *69*, 1061–1065.

Tolin, D.F. (1999). *A revised meta-analysis of psychosocial treatments for PTSD*. Poster presented at AABT, Toronto, November 11–14, 1999.

* Turner, R. M. (2000). Naturalistic evaluation of dialectical behavior therapy-oriented treatment for borderline personality disorder. *Cognitive and Behavioral Practice*, *7*, 413–419.

* Verheul, R., van den Bosch, L. M. C., Koeter, W. J., De Ridder, M. A. J., Stijnen, T., & van den Brink, W. (2003). Dialectical behaviour therapy for women with borderline personality disorder. *British Journal of Psychiatry*, *182*, 135–140.

Wells, A. (2000). *Emotional disorders and metacognition: Innovative cognitive therapy*. Chichester, UK: Wiley.

Wells, A. (2006). Personal communication, July 25.

Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition*. Stanford: Stanford University Press.

* Woods, D. W., Wetterneck, C. T., & Flessner, C. A. (2006). A controlled evaluation of acceptance and commitment therapy plus habit reversal for trichotillomania. *Behaviour Research and Therapy*, *44*, 639–656.

* Zettle, R. D. (2003). Acceptance and commitment therapy (ACT) vs. systematic desensitization in treatment of mathematics anxiety. *The Psychological Record*, *53*, 197–215.

* Zettle, R. D., & Hayes, S. C. (1986). Dysfunctional control by client verbal behavior: The context of reason-giving. *The Analysis of Verbal Behavior*, *4*, 30–38.

* Zettle, R. D., & Rains, J. C. (1989). Group cognitive and contextual therapies in treatment of depression. *Journal of Clinical Psychology*, *45*, 436–445.