Journal of Consulting and Clinical Psychology
2007, Vol. 75, No. 2, 232–243

0022-006X/07/$12.00 DOI: 10.1037/0022-006X.75.2.232

Benchmarks for Psychotherapy Efficacy in Adult Major Depression

Takuya Minami
University of Utah

Bruce E. Wampold and Ronald C. Serlin
University of Wisconsin—Madison

John C. Kircher
University of Utah

George S. (Jeb) Brown
Center for Clinical Informatics

This study estimates pretreatment–posttreatment effect size benchmarks for the treatment of major depression in adults that may be useful in evaluating psychotherapy effectiveness in clinical practice. Treatment efficacy benchmarks for major depression were derived for 3 different types of outcome measures: the Hamilton Rating Scale for Depression (M. A. Hamilton, 1960, 1967), the Beck Depression Inventory (A. T. Beck, 1978; A. T. Beck & R. A. Steer, 1987), and an aggregation of low reactivity–low specificity measures. These benchmarks were further refined for 3 conditions: treatment completers, intent-to-treat samples, and natural history (wait-list) conditions. The study confirmed significant effects of outcome measure reactivity and specificity on the pretreatment–posttreatment effect sizes. The authors provide practical guidance in using these benchmarks to assess treatment effectiveness in clinical settings.

Keywords: effectiveness, psychotherapy, outcome, benchmarking, depression

Although the efficacy of psychotherapy for adult depression has clearly been established (e.g., Lambert & Ogles, 2004; Wampold, 2001), there has been a consistent concern in the field as to whether or not clients treated in clinical settings receive the benefits demonstrated in clinical trials (i.e., effectiveness of treatment; Barlow, 1981; Cohen, 1965; Goldfried & Wolfe, 1998; Luborsky, 1972; Seligman, 1995; Strupp, 1989). Despite some evidence from clinical trials suggesting that efficacy corresponds to effectiveness (e.g., Shadish, Matt, Navarro, & Phillips, 2000; Shadish et al., 1997), there are few outcome data from clinical settings to suggest that treatments in these settings (i.e., treatment as usual; TAU) attain the benefits observed in clinical trials.

Treatment efficacy is often gauged by comparing treatment with no treatment (i.e., wait-list control groups). This strategy is often precluded in clinical settings, however, because control groups (i.e., no-treatment controls) rarely exist in naturalistic settings for practical and ethical reasons. Thus, it is often unclear as to whether the effectiveness of TAU is significantly better than the natural history of the disorder or is as effective as treatments provided in clinical trials.

One method to assess effectiveness in clinical settings is *benchmarking*, a strategy that allows for comparison of outcome data obtained from clinical settings (i.e., TAU) against a reliable outcome standard observed in clinical trials. In the area of childhood depression, Weersing and Weisz (2002) conducted a study in which they compared the outcome data in six community mental health centers in the Los Angeles area against a clinical trials benchmark. Contrary to other benchmarking studies that selected a single clinical trial to serve as a benchmark (e.g., Merrill, Tolbert, & Wade, 2003; Wade, Treat, & Stuart, 1998) Weersing and Weisz conducted a meta-analysis of 13 cognitive–behavioral therapy clinical trials, aggregating the effect sizes to obtain "a research standard of care for comparison—creating a best practice benchmark from a review of the entire youth depression treatment literature" (p. 300).

One factor that Weersing and Weisz (2002) did not explicitly pursue as a potential methodological issue in conducting benchmarking studies was the reactivity and specificity of outcome measures, which have repeatedly been shown to significantly affect the effect size estimates of treatment outcomes (Lambert & Bergin, 1994; Lambert, Hatch, Kingston, & Edwards, 1986; Robinson, Berman, & Neimeyer, 1990; Shadish et al., 1993, 1997, 2000; Shapiro & Shapiro, 1982; Smith, Glass, & Miller, 1980). *Reactivity* is generally concerned with the sensitivity of the measure produced by the rater of the outcome—notably, an observer (either the treating clinician or an independent rater) or the client. *Specificity*, on the other hand, refers to the extent to which the outcome measures assess targeted symptoms of a particular disor-

Takuya Minami and John C. Kircher, Department of Educational Psychology, University of Utah; Bruce E. Wampold, Department of Counseling Psychology, University of Wisconsin—Madison; Ronald C. Serlin, Department of Educational Psychology, University of Wisconsin—Madison; George S. (Jeb) Brown, Center for Clinical Informatics, Salt Lake City, Utah.

Part of this article was based on a doctoral dissertation in partial fulfillment of the requirements for a doctorate in counseling psychology from the University of Wisconsin—Madison, completed by Takuya Minami under the guidance of Bruce E. Wampold and Ronald C. Serlin. Partial funding for this study was provided by the Department of Counseling Psychology, University of Wisconsin—Madison as a doctoral research award to Takuya Minami. We thank Jason A. Seidel for his critique of earlier versions of this article.

Correspondence concerning this article should be addressed to Takuya Minami, Department of Educational Psychology, University of Utah, 1705 East Campus Center Drive, Room 327, Salt Lake City, UT 84112. E-mail: takuya.minami@ed.utah.edu

der rather than global functioning. For example, the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960, 1967) is an outcome measure of specific symptoms of depression (i.e., high specificity) often rated by an observer (i.e., high reactivity). The Global Assessment of Functioning (GAF; American Psychiatric Association, 1994) is also rated by an observer (i.e., high reactivity) but measures global functioning (i.e., low specificity). The Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), on the other hand, is a client self-report measure (i.e., low reactivity) that measures targeted depressive symptoms (i.e., high specificity). Finally, measures such as the Global Severity Index of the Symptom Check List–90—Revised (SCL-90–R; Derogatis, 1983; Derogatis, Rickels, & Rock, 1976) are client self-report measures (i.e., low reactivity) that assess global functioning (i.e., low specificity). Based on previous research, it is suspected that aggregated benchmarks could substantially differ in their effect size based on the reactivity and/or specificity of the outcome measures that were aggregated.

The evidence leading to this hypothesis is substantial. From as early on as the seminal meta-analysis of the efficacy of psychotherapy by Smith et al. (1980; see also Smith & Glass, 1977), measurement reactivity and specificity were found to significantly affect the treatment effect size—in particular, measures higher in specificity and reactivity were found to be associated with larger effect sizes. Shortly thereafter, this finding was replicated by Shapiro and Shapiro (1982). A more recent meta-analysis of family and marital therapies by Shadish et al. (1993) also confirmed that more reactive and specific outcome measures produced larger effects than less reactive and specific measures. Specific to reactivity, a meta-analysis by Lambert, Hatch, Kingston, and Edwards (1986) found that the HRSD showed significantly more change than the BDI or the Zung Self-Rating Scale (Zung, 1965). Therefore, it is reasonable to suspect that aggregated effect sizes must consider measurement reactivity and specificity.

Another factor that potentially influences the benchmarks is client attrition. Unlike clinical trials, clinical settings rarely specify an a priori duration of treatment. Therefore, treatment effects measured in clinical settings often include all clients seeking treatment by comparing the initial measurement (i.e., pretreatment) to the final measurement assessed prior to the end of the last treatment episode. This method corresponds to the use of intent-to-treat (ITT) samples in efficacy research, which include all clients who were initially randomized into conditions. In contrast, clinical trial outcomes are often calculated based only on those who completed the treatment protocol (completer samples). Therefore, assessing the effectiveness of TAU data may require efficacy benchmarks derived from ITT samples. Furthermore, because ITT samples include those who do not continue treatment for the specified duration, it is likely that such samples produce smaller effects than completer samples.

The purpose of the present study was to construct psychotherapy treatment outcome benchmarks for the outpatient treatment of adult major depression, by meta-analytically aggregating outcomes from clinical trials that were methodologically stringent, and to assess how reactivity and specificity of the outcome measures affect the benchmarks. In addition, aggregated effect sizes were compared between completer samples and ITT samples to determine whether or not excluding dropouts in the analysis would lead to larger effect sizes. As well, if effect size estimates were found to be heterogeneous, various factors that might moderate the estimates were examined, including initial severity, type of treatment, and modality (group vs. individual). Benchmarks were also constructed for the natural course of major depression by examining no-treatment controls, so that outcomes in clinical settings could be compared with no-treatment as well as treatment efficacy benchmarks. Finally, critical values for the benchmarks were calculated for different sample sizes to provide ranges for clinically nonsignificant differences between TAU outcomes and the benchmarks.

Method

Efficacy Benchmarks for Major Depression

Two methods were used to identify psychotherapy clinical trials for adult major depression. First, independent clinical trials of depression that were included in meta-analytic reviews of psychotherapy were compiled (viz., Ahn & Wampold, 2001; Areán & Cook, 2002; Bower, Rowland, & Hardy, 2003; Crits-Christoph, 1992; DeRubeis, Gelfand, Tang, & Simons, 1999; Dobson, 1989; Gaffan, Tsaousis, & Kemp-Wheeler, 1995; Gloaguen, Cottraux, Cucherat, & Blackburn, 1998; McCullough, 1999; Posternak & Miller, 2001; Robinson et al., 1990; Svartberg & Stiles, 1991; Thase et al., 1997; Wampold et al., 1997; Wampold, Minami, Baskin, & Tierney, 2002; Westen & Morrison, 2001). Second, independent clinical trials between the years 1995 and 2003 were manually searched in PsycINFO using the search term *((therapy OR counseling OR psychotherapy) AND LA=ENGLISH AND (depressi* NOT (manic OR mania OR bipolar)))*, which returned 11,779 entries. Combined, these two methods resulted in 383 published articles assessing the efficacy of psychotherapy treatments for depression. Next, the 383 articles were screened to exclude duplicate data (i.e., multiple publications using the same clinical trials data) and studies in which the participants were not adult outpatients (i.e., ages 18 and older), leaving 224 independent clinical trials. Then, the clinical trials were thoroughly evaluated based on the following inclusion criteria: (a) Clients had clinically significant symptoms of unipolar major depression, (b) clients were randomly allocated to treatment condition, (c) clients received bona fide outpatient psychotherapy (for the treatment benchmark), (d) studies reported sufficient data to calculate effect sizes, and (e) clients were not concurrently on medication or placebo. In addition, a subset of the studies that reported results with ITT samples were selected to create benchmarks that would approximate TAU data.

Unipolar major depression was determined by (a) diagnosis of major depression using *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) criteria (e.g., *DSM,* 4th ed.; *DSM–IV*; American Psychiatric Association, 1994), (b) formal diagnostic interview with instruments such as the Structured Clinical Interview for *DSM* (Spitzer & Williams, 1984; Spitzer, Williams, Gibbon, & First, 1989, 1995), or (c) pretest scores on measures of depression commonly corresponding to a *DSM* diagnosis of major depression (e.g., a BDI score of 14 or above). Studies that reported fulfilling one of the above criteria but not another were excluded (e.g., studies with clients scoring 14 or above on the BDI but clearly stating that some clients had minor rather than major depression). We also excluded studies with a primary treatment focus on

substance use, personality disorder, or depressive symptoms secondary to medical conditions. Randomization was included as a criterion to filter out clinical trials that allowed clients to choose their preferred treatment. Therefore, in studies that used both a randomized and a client preference arm, only outcome data from the randomized arm were included.

The criteria for evaluating a psychotherapy treatment as bona fide were adapted from Wampold et al. (1997). First, treatment was conducted by a clinician with at least a master's degree in a relevant field who met face-to-face with clients. Second, treatment was tailored to the individual receiving the treatment. Third, treatment met at least two of the following criteria: (a) a citation was given to an established psychotherapy approach (e.g., Rogers, 1951), (b) a description of the therapy with a reference to a psychological process (e.g., positive reinforcement) was provided in the study, (c) a manual was used to guide the administration of the therapy (e.g., Beck, Rush, Shaw, & Emery, 1979), and/or (d) citations were provided for the identified active ingredients of the treatment.

Of the 224 independent clinical trials, only 35 studies met the full inclusion criteria. Specifically, 100 studies were excluded because the sample did not meet criteria for major depression; 31 studies were excluded because clients were not randomized into conditions; 8 studies were excluded because their treatment condition was not deemed bona fide; 32 studies did not report sufficient data to calculate an effect size; and 18 studies simultaneously prescribed either antidepressants or placebos. Accurate demographic data for the clients in the 35 studies could not be determined, as information provided in the original articles was sparse; however, it was estimated that over 70% were female and over 90% were Caucasian. The studies included in the benchmarks are marked in the References by an asterisk.

The 35 studies were further investigated to see whether or not they also reported results of their ITT sample. In studies in which the researchers were unable to assess the outcomes of clients who prematurely terminated, studies were considered ITT provided that missing endpoints were substituted with the last observation and were carried forward. Eleven of the 35 studies reported results for their ITT sample (marked with a double asterisk in the References). In addition, the following variables were coded for each study in order to conduct a moderator analysis: initial severity, treatment type (i.e., cognitive/behavioral/cognitive–behavioral vs. other), modality (i.e., individual vs. group), weeks in treatment, and sample size in the studies.

Natural History Benchmarks for Major Depression

Identification of studies that would provide data for the natural history of depression between pre- and posttreatment was facilitated by a meta-analysis by Posternak and Miller (2001) that estimated wait-list control group effect sizes. However, because their inclusion criteria were slightly different than ours, we reexamined all of the studies that they included. With an additional search using PsycINFO (as described above), a total of 11 studies met our criteria for inclusion. These studies are marked with a dagger in the References.

Categorization of Outcome Measures

The measures used in the original studies were categorized based on reactivity and specificity. Consistent with previous research, measures were considered high on reactivity if they were assessed by an independent clinician and low on reactivity if clients provided self-report data. In addition, the specificity of the measures was assessed: Measures that focused specifically on symptoms of depression were categorized as high specificity, and those focusing on broader symptoms and global functioning as low specificity. Therefore, four categories of outcome measures were initially created: (a) high reactivity–high specificity (HR-HS), (b) high reactivity–low specificity (HR-LS), (c) low reactivity–high specificity (LR-HS), and (d) low reactivity–low specificity (LR-LS). Here, it is noted that all studies that included clinician-rated measures used an independent rater (i.e., the rating was not conducted by the treating clinician). However, effect sizes for the HR-LS benchmarks for both treatment and natural history were excluded because of low number of studies and resulting total sample size. In addition, as all of the studies that included measures classified in the HR-HS category used the HRSD, and similarly, as all studies that included LR-HS measures used the BDI, the benchmarks for the respective outcome measure categories were calculated using only data from the HRSD and the BDI. For the LR-LS benchmark, measures were excluded if they (a) did not assess global functioning or symptoms broader than those that are specific to major depression, (b) were designed to assess specific symptoms of other *DSM–IV* diagnoses (e.g., Hamilton Rating Scale for Anxiety; Hamilton, 1959), (c) were designed for specific types of therapies (e.g., Automatic Thoughts Questionnaire; Hollon & Kendall, 1980), and (d) were deemed irrelevant to psychotherapy outcome (i.e., Religious Behavior Scale; Johnson, DeVries, Ridley, Pettorini, & Peterson, 1994). By far the most common measure included in the LR-LS benchmarks was the SCL-90–R or its Global Severity Index. In summary, treatment efficacy and natural history benchmarks were calculated under three conditions (i.e., ITT, completers, wait-list) and three outcome measure categories (i.e., HRSD, BDI, LR-LS), resulting in a total of nine benchmarks.

Calculation of Benchmarks

Benchmark calculations were conducted following standard meta-analytic procedures developed by Hedges and Olkin (1985) and Becker (1988). Broadly, for all benchmarks, two steps were taken for the calculation. The first step involved aggregating the pretreatment–posttreatment data within each study i that used an outcome measure category j to obtain a single pretreatment–posttreatment effect size estimate d_{ij}. The second step involved aggregating each of the effect size estimates to obtain a single pretreatment–posttreatment effect benchmark d_{+J} for each outcome measure category. For studies that involved more than one treatment, the results of all bona fide treatments were aggregated for the treatment efficacy benchmarks because (a) the primary purpose of this study was not to compare differential efficacies among different bona fide psychotherapies but to create representative benchmarks for bona fide clinical trials in general, and (b) studies comparing different psychotherapy treatments for adult depression have resulted in equivalent efficacy (e.g., Wampold et

al., 1997; Luborsky et al., 2002); however, a moderator analysis was conducted to confirm equivalence among different types of treatment. When studies reported more than one outcome measure that could be classified to fall within the LR-LS category, the unbiased estimators d_{ij} for the outcome measures were aggregated within the studies before aggregating across studies, following Gleser and Olkin (1994).

Moderator Analysis

Five moderators—initial severity, treatment type (cognitive/behavioral/cognitive–behavioral vs. other), modality (individual vs. group), weeks in treatment, and sample size—were tested as potential moderators. Initial severity is well-known to affect estimates of effect sizes, in that more distressed clients experience more gains by the end of treatment (Garfield, 1986; Lambert, 2001). Psychotherapy researchers are interested in the effects of treatment type and modality to identify the most efficacious treatments, and these may be important variables in establishing any benchmark. Number of weeks in treatment was included, as the dose–effect relationship in psychotherapy is well documented (e.g., Howard, Kopta, Krause, & Orlinsky, 1986). Finally, sample size of the individual studies was included, as studies with smaller sample sizes may have larger effect sizes based on statistical power and publication bias toward studies with statistical significance (Quintana & Minami, 2006).

Moderator analysis was conducted using the multiple regression analogue mixed effects model as developed by Hedges and Pigott (2004). Specifically, the mixed effects model incorporates both the sampling error v_{ij} and the residual variance component τ^2 of the effect size estimates d_{ij}, thus modeling the variance as $vij * = vij + \tau^2$. Whereas v_{ij} values are known, τ^2 is estimated by the equation

$$\hat{\tau}^2 = \frac{Q_E - (k - p - 1)}{a}, \tag{1}$$

where k is the number of effect size estimates and p is the number of moderators. Q_E, the test of goodness of fit of the fixed-effects regression model, is defined by the equation

$$Q_E = \mathbf{d}'[\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{d}, \tag{2}$$

where $\mathbf{d} = d_{1j}, \ldots, d_{kj}$ for the k studies measured by respective outcome measure j, $\mathbf{V} = \text{Diag}(v_{1j}, \ldots, v_{kj})$, and $\mathbf{X}$ is a $k \times (p + 1)$ design matrix with a vector of ones in the first column and moderators in the other columns. The constant a is defined as follows:

$$a = \sum_{i=1}^{k} v_{ij}^{-1} - \text{tr}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-2}\mathbf{X}]. \tag{3}$$

When $\mathbf{V}^* = \text{Diag}(v_{1j} + \tau^2, \ldots, v_k + \tau^2)$, the regression coefficients are estimated by

$$\hat{B}^* = [\mathbf{X}'(\mathbf{V}^*)^{-1}\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{V}^*)^{-1}\mathbf{d}, \tag{4}$$

which has a covariance matrix

$$\Sigma^* = [\mathbf{X}'(\mathbf{V}^*)^{-1}\mathbf{X}]^{-1}. \tag{5}$$

Comparisons Among Benchmarks

Comparisons between completer samples and ITT samples. With studies that reported results of both completer and ITT samples, effect size estimates were compared to determine the extent to which excluding dropouts influenced the effect size of the treatment. It was hypothesized that the exclusion of dropouts would result in significantly larger pretreatment–posttreatment effect sizes when compared against the outcomes of ITT samples.

Effect of reactivity and specificity. For the efficacy benchmarks, two comparisons were conducted using the studies with completer samples. On the basis of previous research, we hypothesized that higher reactivity and specificity would result in larger effect sizes. Specifically, the HRSD efficacy benchmark was expected to be significantly larger than the BDI benchmark (Comparison 1) because of differences in reactivity. The BDI efficacy benchmark, on the other hand, was expected to be significantly larger than the LR-LS efficacy benchmark because of differences in specificity (Comparison 2). These comparisons were conducted using eight studies with completer data that used all three outcome measure categories to avoid confound due to possible differences in sampled populations and design characteristics. Parallel comparisons using the natural history benchmarks were not conducted because of the low number of studies and resulting total sample size.

Calculation of Critical Values for Use in Clinical Settings

In order to statistically claim that any treatment effect size estimate obtained from clinical settings is comparable to treatment efficacy benchmarks or is greater than natural history benchmarks, the estimate must reach certain critical values, which are dependent on the sample size of the clinical setting data. On the basis of Cohen's (1988, 1992) assertion, a minimum effect size of $d_{\min} = 0.2$ was adopted as the criterion for clinically significant differences between benchmarks and the treatment effect size estimates for a range-null hypothesis test. That is, if the true effect size was within 0.2 standard deviations of the efficacy benchmark, it was determined to be clinically equivalent to the clinical trials. This conservative criterion was selected because the aim was to create a criterion that would indicate that the clinical settings effect size estimate of interest is clinically equivalent to that of the best clinical trials. On the other hand, if the true effect size was within 0.2 standard deviations of the natural history benchmark, one could conclude that the delivered treatment had no clinically meaningful effect. Therefore, the critical value for the treatment efficacy benchmark should allow the conclusion that the true treatment effect size is no smaller than the value that is 0.2 standard deviations below the efficacy benchmark, while maintaining a Type I error rate of .05. The critical value for the natural history benchmark, on the other hand, should allow the conclusion that the true treatment effect size is larger than the value that is 0.2 standard deviations above the natural history benchmark, again with a Type I error rate of .05.

Calculations of these critical values were conducted using the range-null hypothesis testing procedure developed by Serlin and Lapsley (1985, 1993; see also Minami, Wampold, Serlin, Kircher, & Brown, in press), as the distribution of these critical values followed a noncentral t distribution. Specifically, when d_{CV} is the

critical effect size value for the clinical settings data with a sample size of N to exceed,

$$d_{CV} = t_\lambda/\sqrt{N}, \quad (6)$$

where t_λ is the 95th percentile t value with a noncentrality parameter

$$\lambda = \sqrt{N}(d_{+J} \pm d_{\min}). \quad (7)$$

Here, as above, d_{+J} is the benchmark and $d_{\min} = 0.2$. The sign between d_{+J} and $d_{\min}$ is based on whether the critical value is for comparison against the treatment efficacy (negative) or natural history (positive). The critical values were calculated for hypothetical clinical sample sizes from 100 to 10,000.

Results

Clinical Trials of Depression Benchmarks

The aggregated treatment efficacy and natural history benchmarks are listed in Table 1. Although precise statistical comparisons are to follow in the next section, the magnitude of the efficacy benchmark was the largest for the HRSD, with effect sizes of $d_+ = 2.249$ (completers) and $d_+ = 2.434$ (ITT). The BDI benchmark resulted in $d_+ = 1.859$ (completers) and $d_+ = 1.706$ (ITT). Finally, the LR-LS measures resulted in the smallest aggregate effect sizes of $d_+ = 0.932$ (completers) and $d_+ = 0.795$ (ITT). The typical duration of the treatments was estimated as 15 weeks for the completer sample and 16 weeks for the ITT sample. With regard to the natural history benchmarks, the benchmark obtained from aggregating the HRSD was similar to that obtained from aggregating the BDI. As the Q statistics for homogeneity indicated that effect size estimates were heterogeneous other than for the BDI and LR-LS natural history benchmarks, the reported benchmarks should not be considered an estimate of a single parameter but rather should be considered as the mean of the effect size estimates (Shadish & Haddock, 1994).

The following five moderators were tested: initial severity, treatment type (i.e., cognitive/behavioral/cognitive–behavioral vs. other), modality (i.e., individual vs. group), weeks in treatment, and sample size in the studies (see Tables 2–5). All BDI benchmarks were moderated by initial severity, where higher severity was related to larger effect size. As consistent with previous research, number of weeks in treatment was significantly and positively related to the HRSD treatment efficacy benchmark (completers). However, weeks in treatment was significantly and negatively related to the LR-LS treatment efficacy benchmark (completers). Treatment type, modality, and sample size did not significantly affect the benchmarks.

Statistical Comparisons of Clinical Trials of Depression Benchmarks

Comparisons between completers and ITT samples. Treatment efficacy effect size differences were analyzed using studies that reported both completers and ITT samples. Table 6 summarizes the analyses under each outcome measure category. Analyses revealed that completer effect size estimates were significantly greater than ITT effect size estimates in all outcome measure categories, demonstrating that the exclusion of clients failing to complete the protocol inflated the obtained estimate of treatment effects. It is noted that as the analyses were within-studies comparisons, the between-studies moderators reported in Tables 2 through 5 do not affect the results.

Effect of reactivity and specificity. Statistical analyses comparing the different benchmarks were conducted by aggregating the eight studies that included all three outcome measure categories. Therefore, although these comparisons do not provide an exact test of the relative differences of the aggregated benchmarks, they nevertheless provide the best estimate of the effect of reac-

Table 1
Aggregated Benchmarks

Measure	K	N	d_+	$\sigma^2_{d(+)}$	$d_{+.975}$	$d_{+.025}$	Q	$p(Q)$
Treatment efficacy benchmarks (ITT)								
HRSD	8	458	2.434	0.009	2.619	2.250	23.26	.002
BDI	11	846	1.706	0.003	1.815	1.596	120.56	<.001
LR-LS	4	489	0.795	0.003	0.894	0.695	26.52	<.001
Treatment efficacy benchmarks (completers)								
HRSD	24	1,107	2.249	0.003	2.363	2.134	135.66	<.001
BDI	29	1,387	1.859	0.002	1.950	1.768	228.24	<.001
LR-LS	11	768	0.932	0.002	1.014	0.851	88.26	<.001
Natural history benchmarks								
HRSD	9	122	0.401	0.010	0.595	0.207	23.49	.003
BDI	11	140	0.371	0.008	0.547	0.196	12.98	.225
LR-LS	5	68	0.149	0.013	0.370	−0.073	1.84	.765

Note. K = number of studies; Q = test of homogeneity; ITT = intent to treat; HRSD = Hamilton Rating Scale for Depression (Hamilton, 1960, 1967); BDI = Beck Depression Inventory (Beck, 1978; Beck & Steer, 1987); LR-LS = aggregate of low reactivity and low specificity outcome measures.

Table 2
Moderator Effects on Treatment Efficacy and Natural History Benchmarks: Severity

Measure	*M*	*SD*	*B*	*SE B*	*p*
			Treatment efficacy benchmarks (ITT)		
HRSD	19.89	4.06	0.023	0.206	.911
BDI	26.95	7.57	0.243	0.071	<.001
LR-LS[a]	—	—	—	—	—
			Treatment efficacy benchmarks (completers)		
HRSD	20.50	4.59	0.056	0.031	.075
BDI	26.49	7.37	0.095	0.040	.017
LR-LS[a]	—	—	—	—	—
			Natural history benchmarks		
HRSD	19.46	4.88	0.008	0.037	.833
BDI	23.14	6.93	0.053	0.026	.040
LR-LS[a]	—	—	—	—	—

Note. ITT = intent to treat; HRSD = Hamilton Rating Scale for Depression (Hamilton, 1960, 1967); BDI = Beck Depression Inventory (Beck, 1978; Beck & Steer, 1987); LR-LS = aggregate of low reactivity and low specificity outcome measures.
[a] Severity was not established because of the composite measure.

Table 4
Moderator Effects on Treatment Efficacy and Natural History Benchmarks: Weeks in Treatment

Measure	*M*	*SD*	*B*	*SE B*	*p*
			Treatment efficacy benchmarks (ITT)		
HRSD	16.49	6.81	0.045	0.119	.702
BDI	16.25	5.29	−0.047	0.052	.359
LR-LS	16.01	2.18	−0.036	0.071	.607
			Treatment efficacy benchmarks (completers)		
HRSD	14.87	5.16	0.053	0.022	.018
BDI	14.75	4.80	−0.006	0.021	.777
LR-LS	15.51	3.76	−0.051	0.023	.030
			Natural history benchmarks		
HRSD	9.70	3.31	0.074	0.067	.273
BDI	9.54	3.19	0.050	0.028	.079
LR-LS	9.01	3.26	0.022	0.038	.569

Note. ITT = intent to treat; HRSD = Hamilton Rating Scale for Depression (Hamilton, 1960, 1967); BDI = Beck Depression Inventory (Beck, 1978; Beck & Steer, 1987); LR-LS = aggregate of low reactivity and low specificity outcome measures.

tivity and specificity. Table 7 summarizes the statistical analyses of the two planned comparisons. Both comparisons were significant in the direction hypothesized, supporting the effect of both reactivity and specificity on the estimated effect size of treatments. It is noted that as these analyses were also within-studies comparisons, the results are not affected by the between-studies moderators reported in Tables 2 through 5.

Critical Values for Use With Clinical Settings Data

Critical values for the treatment efficacy and natural history benchmarks using the ITT samples of the HRSD, BDI, and LR-LS are presented in Figures 1, 2, and 3, respectively. Here it is noted that as these calculated values are not adjusted based on the moderator variables, adjustments may be necessary depending on the clinical settings data that one intends to benchmark. For example, when using the BDI benchmarks, severity can be matched by two possible methods. The first method is to adjust the benchmark based on the coefficients and means reported in Tables 2 through 5. Let us say that the mean initial severity using the BDI was 25.95 for the clinical settings data, creating a mean difference of 1.00 between the treatment efficacy benchmark (ITT) and the data. Therefore, the adjusted benchmark would be 0.243 lower than the benchmark, notably, 1.463. The adjusted critical values

Table 3
Moderator Effects on Treatment Efficacy: Treatment Type and Modality

	Treatment type				Modality			
Measure	%[a]	*B*	*SE B*	*p*	%[b]	*B*	*SE B*	*p*
				Treatment efficacy benchmarks (ITT)				
HRSD	82.97	0.795	0.685	.246	92.79	0.677	2.930	.817
BDI	71.16	0.270	0.312	.387	96.10	−2.565	1.561	.100
LR-LS	54.21	−0.120	0.345	.727	100.00		—[c]	
				Treatment efficacy benchmarks (completers)				
HRSD	85.91	0.051	0.332	.879	59.53	0.437	0.259	.091
BDI	72.07	0.152	0.272	.578	65.22	0.416	0.246	.090
LR-LS	58.41	−0.158	0.194	.415	70.14	−0.035	0.202	.864

Note. Moderator effects of treatment type and modality are not applicable for natural history benchmarks. ITT = intent to treat; HRSD = Hamilton Rating Scale for Depression (Hamilton, 1960, 1967); BDI = Beck Depression Inventory (Beck, 1978; Beck & Steer, 1987); LR-LS = aggregate of low reactivity and low specificity outcome measures.
[a] Percent cognitive/cognitive–behavioral/behavioral. [b]Percent individual therapy. [c]All studies were individual therapy.

Table 5
Moderator Effects on Treatment Efficacy and Natural History Benchmarks: Sample Size

Measure	*M*	*SD*	*B*	*SE B*	*p*
		Treatment efficacy benchmarks (ITT)			
HRSD	45.80	37.99	−0.001	0.008	.899
BDI	56.40	44.04	−0.003	0.004	.460
LR-LS	69.57	42.81	0.000	0.005	.975
		Treatment efficacy benchmarks (completers)			
HRSD	36.90	35.31	−0.002	0.003	.621
BDI	36.47	32.06	−0.004	0.004	.346
LR-LS	47.93	30.35	0.001	0.003	.864
		Natural history benchmarks			
HRSD	13.56	4.36	−0.001	0.051	.984
BDI	12.73	4.54	−0.020	0.021	.349
LR-LS	13.60	5.64	−0.026	0.025	.304

Note. ITT = intent to treat; HRSD = Hamilton Rating Scale for Depression; BDI = Beck Depression Inventory; LR-LS = aggregate of low reactivity and low specificity outcome measures.

would then be calculated by deriving the noncentrality parameters as specified in Equation 7. An alternative method would be to use a subsample in the clinical setting that approximates the severity of the benchmarks estimated in this meta-analysis.

An example of benchmarking is illustrated using the HRSD, assuming that the average number of weeks in treatment is equivalent to our benchmark. The HRSD treatment efficacy benchmark (ITT) is 2.434. As the sample size of the clinical settings data affects the critical values, if the clinical setting sample size is relatively small (e.g., 100), then the observed effect must be considerably larger than the efficacy benchmark minus the 0.2 band (i.e., 2.234). Thus, for 100 clients, the observed effect must be 2.581 to reject the null hypothesis that it is smaller than the 0.2 effect size unit below the efficacy benchmark. That is, if the obtained pretreatment–posttreatment effect for the HRSD for these 100 clients is larger than 2.581, it can be concluded that the treatment is clinically equivalent to treatments in clinical trials. As the clinical settings data sample size increases, the critical values approach the 0.2 band. For example, if the clinical sample is 2,000, the critical value to exceed to claim clinical equivalence with the clinical trials is 2.305.

To say that the treatment in practice is superior to no treatment by a meaningful difference (again, greater than the 0.2 effect size), the obtained effect will need to be significantly greater than the natural history benchmark plus 0.2 (i.e., 0.601). Therefore, again, for a small sample size (e.g., 100), the observed pretreatment–posttreatment effect size on the HRSD must be greater than the critical value 0.791 to conclude that the treatment is clinically superior to no treatment; for a sample size of 2,000, the effect size needs to be greater than 0.641 to conclude that the treatment is clinically superior to no treatment. It also is possible that a treatment produces an effect that is clinically superior to no treatment, but not clinically equivalent to the efficacy benchmark (e.g., an observed pretreatment–posttreatment HRSD effect size of 1.500 for a sample of 800). In summary, if the clinical setting data exceed the efficacy critical value, this provides evidence that the treatment is clinically equivalent to treatment in controlled clinical trials. If the data do not exceed the efficacy critical value but exceed the natural history critical value, this indicates that although the treatment has a clinical effect beyond natural remission, it is not clinically equivalent to the best clinical trials. If the data do not exceed the natural history critical value, this indicates that the treatment does not have any clinically meaningful effect.

Discussion

The present study provides benchmarks of psychotherapy efficacy for adult depression treatment that can readily be used to assess treatment effectiveness in clinical settings. To ensure the reliability and validity of the psychotherapy benchmarks, stringent inclusion and exclusion criteria resulted in inclusion of only 35 published clinical trials out of 383 psychotherapy studies reviewed (224 independent trials). Furthermore, only 11 studies met the most rigorous criterion of reporting ITT samples, and only 8 studies included at least three categories of outcome measures based on reactivity and specificity that affect the effect size estimates. Analyses between completers and ITT samples supported the hypothesis that including only those participants who completed the specified number of sessions inflated the observed pretreatment–posttreatment effect sizes. As such, naturalistic clinical settings data (i.e., that do not exclude data based on treatment "completion") should be compared with the ITT benchmarks; benchmarks computed using completers should be used when clinical setting "completers" are extracted for analysis.

Consistent with the literature, the present study also confirmed the significant impact of measurement reactivity and specificity on the efficacy benchmarks. Specifically, with regards to reactivity, the HRSD benchmark was significantly larger than that of the BDI. With regards to specificity, the BDI benchmark was significantly

Table 6
Comparisons Between Completers and ITT Samples

Category	*K*	N_C	N_{ITT}	d_C	d_{ITT}	d_-	$\sigma^2_{d(-)}$	*CV*	*p*	$CI_{.95}$
HRSD	6	303	391	3.223	2.548	0.675	0.0138	0.230	<.001	$d_- \geq 0.482$
BDI	8	605	747	2.023	1.725	0.297	0.0003	0.033	<.001	$d_- \geq 0.270$
LR-LS	3	356	457	0.974	0.780	0.194	0.0001	0.023	<.001	$d_- \geq 0.175$

Note. The effect sizes d_C and d_{ITT} are those of completers and intent-to-treat (ITT) samples, respectively. HRSD = HRSD = Hamilton Rating Scale for Depression (Hamilton, 1960, 1967); BDI = Beck Depression Inventory (Beck, 1978; Beck & Steer, 1987); LR-LS = aggregate of low reactivity and low specificity outcome measures. *CV* = critical value at $\alpha = .05$; CI_{95} = confidence interval of the difference at 95%.

Table 7
Effect of Reactivity and Specificity

Comparison	*K*	*N*	d_{1+}	d_{2+}	d_{-}	$\sigma^2_{d(-)}$	*CV*	*p*	$CI_{.95}$
1. HRSD vs. BDI	8	441	2.181	1.831	0.350	0.007	0.091	>.001	$d_{-} \geq 0.212$
2. BDI vs. LR-LS	8	441	1.831	0.969	0.862	0.008	0.125	>.001	$d_{-} \geq 0.715$

Note. The effect sizes d_{1+} and d_{2+} are in the order of the specified outcome measure categories. HRSD = Hamilton Rating Scale for Depression (Hamilton, 1960, 1967); BDI = Beck Depression Inventory (Beck, 1978; Beck & Steer, 1987); LR-LS = aggregate of low reactivity and low specificity outcome measures. *CV* = critical value at $\alpha = .05$; CI_{95} = confidence interval of the difference at 95%.

larger than that of the LR-LS measures. This indicates that benchmarking should be conducted by matching the outcome measures with regard to both reactivity and specificity so that the comparisons yield valid conclusions.

Several variables moderated the aggregated benchmarks. In general, there are two ways to accommodate these moderators in a benchmarking study. First, the benchmarks can be adjusted using the coefficients reported and the benchmarks adjusted accordingly as described previously. Second, a subsample of clinical data that matches the parameters of the meta-analysis can be used. For example, cases could be selected in such a way that the dosage and severity match the average values of the studies composing the meta-analysis reported here.

There are a number of limitations to this study that must be considered when drawing inferences or applying the results to clinical settings. First is the low number of clinical trials that were included in this study, especially for the benchmarks using the ITT samples. Although relaxing some of our inclusion criteria would have greatly increased the number of studies to be included, this would have increased heterogeneity among the studies, possibly introducing more error variance and thus limiting our conclusions and potential applications.

Second, for most benchmarks, the effect size estimates that were aggregated were heterogeneous. Although statistical heterogeneity does not preclude aggregating effect size estimates (Shadish & Haddock, 1994), use and interpretation of the benchmarks warrant caution, as the benchmarks cannot be considered an estimate of a single population parameter.

Third, it is important to point out that because the critical values to be exceeded to attain clinical equivalence or claim treatment effect rise sharply as sample size of the clinical settings data become smaller (i.e., fewer than 100), benchmarking would be impractical for use with sample sizes below 100. In such cases, although clinical settings could compare the observed effect size against the treatment efficacy benchmark minus d_{min} (and conversely, natural history benchmark plus d_{min}), the results would only pertain to the current set of data and would not be generalizable. For example, even if data

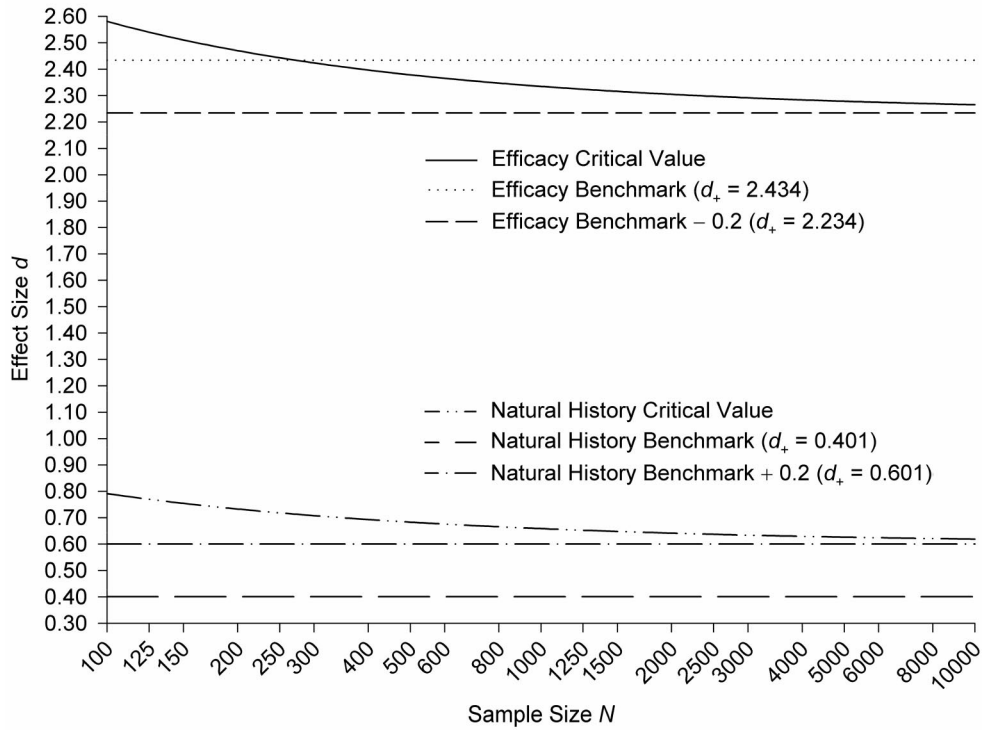

Figure 1. Hamilton Rating Scale for Depression effect size critical values by clinical data sample size.

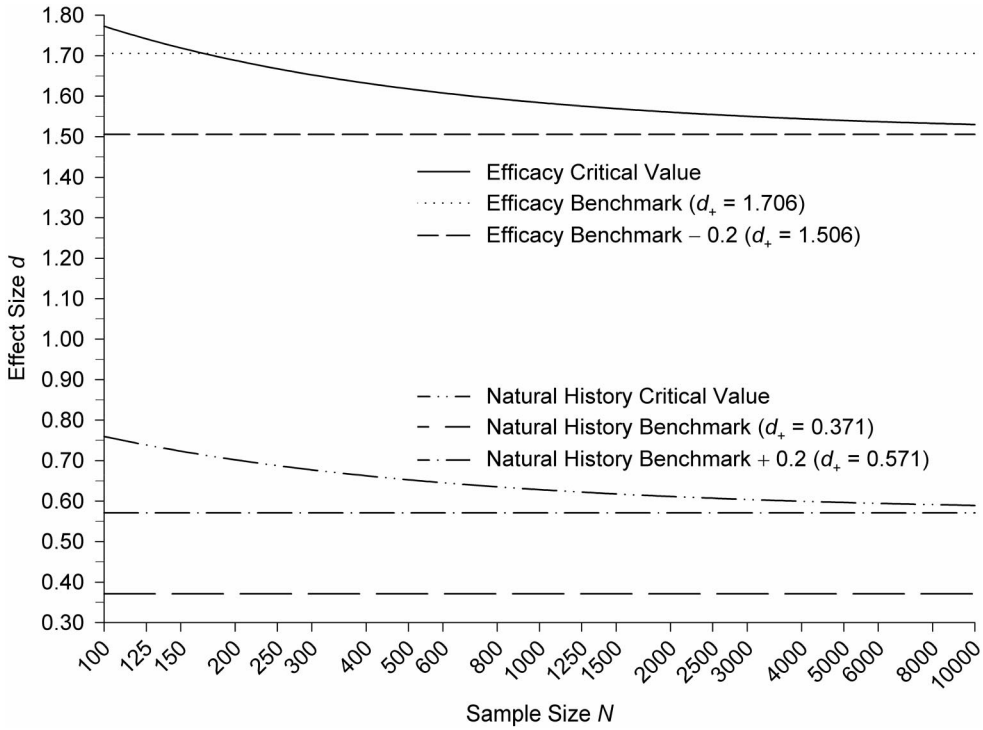

Figure 2. Beck Depression Inventory effect size critical values by clinical data sample size.

exceed the efficacy benchmark minus d_{min} for 50 clients, the only conclusion that could be drawn from this result is that for these 50 clients, the treatment was as clinically effective as the clinical trials.

As a final but critical caveat, it is important to note that there are numerous differences between clinical trials and naturalistic settings that could render simplistic numerical comparisons problematic. In clinical trials, the clients are selected through several

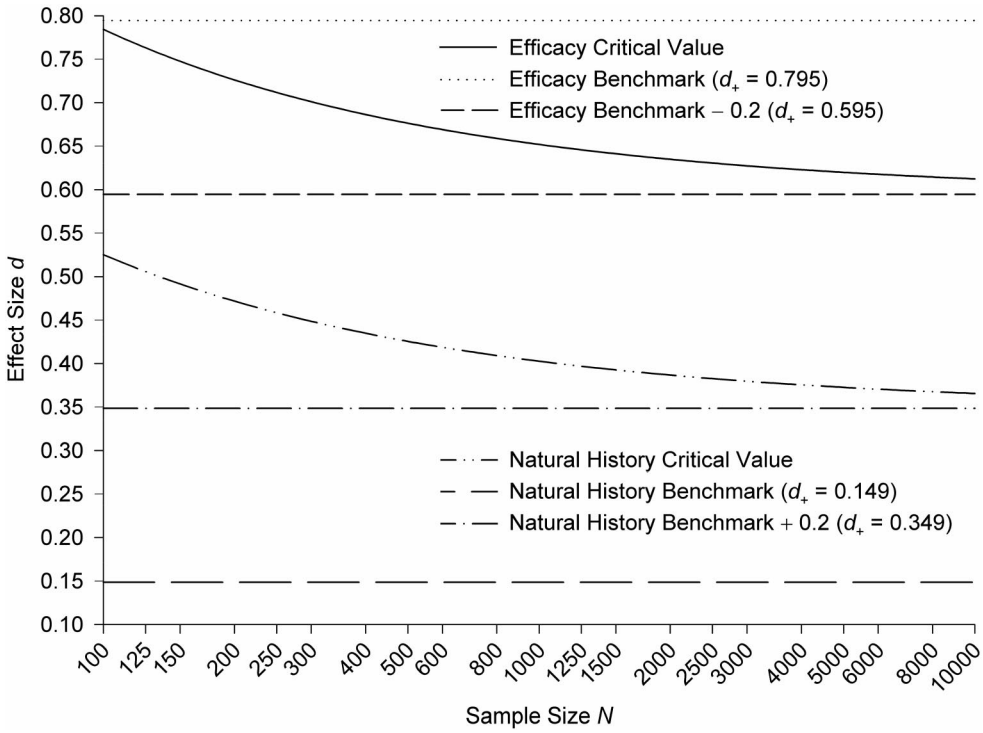

Figure 3. Low reactivity–low specificity outcome measures effect size critical values by clinical data sample size.

inclusion and exclusion criteria; typically, individuals with suicidal ideation, co-occurring substance abuse, personality disorder, manic symptoms, and/or psychosis are excluded (Westen & Morrison, 2001). In addition, whereas therapists in clinical trials are often selected for their expertise, trained, supervised, and/or held to a certain level of clinical expertise (e.g., Rounsaville, O'Malley, Foley, & Weissman, 1988), data in clinical settings typically are obtained from a wider range of therapists. Given that the variability in outcomes is significantly more attributable to therapists than treatments (e.g., Kim, Wampold, & Bolt, 2006), it is reasonable to conclude that clinical settings are disadvantaged to produce treatment effects when compared with clinical trials. However, as the conditions in which clinical trials are conducted provide the assurance that their observed efficacy provides the best standard, it is also reasonable to consider these benchmarks as the numeric criteria to strive for.

References

References marked with an asterisk (or asterisks) and/or a dagger indicate studies included in the meta-analysis.

Ahn, H., & Wampold, B. E. (2001). Where oh where are the specific ingredients? A meta-analysis of component studies in counseling and psychotherapy. *Journal of Counseling Psychology, 48,* 251–257.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Areán, P. A., & Cook, B. L. (2002). Psychotherapy and combined psychotherapy/pharmacotherapy for late life depression. *Biological Psychiatry, 52,* 293–303.

*†Areán, P. A., Perri, M. G., Nezu, A. M., Schein, R. L., Christopher, F., & Joseph, T. X. (1993). Comparative effectiveness of social problem-solving therapy and reminiscence therapy as treatments for depression in older adults. *Journal of Consulting and Clinical Psychology, 61,* 1003–1010.

Barlow, D. H. (1981). On the relation of clinical research to clinical practice: Current issues. *Journal of Consulting and Clinical Psychology, 49,* 147–155.

Beck, A. T. (1978). *Depression inventory.* Philadelphia: Center for Cognitive Therapy.

*Beck, A. T., Hollon, S. D., Young, J. E., Bedrosian, R. C., & Budenz, D. (1985). Treatment of depression with cognitive therapy and amitriptyline. *Archives of General Psychiatry, 42,* 142–148.

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression.* New York: Guilford Press.

Beck, A. T., & Steer, R. A. (1987). *Beck Depression Inventory manual.* San Antonio, TX: Harcourt Brace Jovanovich.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41,* 257–278.

*Beutler, L. E., Engle, D., Mohr, D., Daldrup, R. J., Bergan, J., Meredith, K., & Merry, W. (1991). Predictors of differential response to cognitive, experiential, and self-directed psychotherapeutic procedures. *Journal of Consulting and Clinical Psychology, 59,* 333–340.

*Blackburn, I.-M., & Moore, R. G. (1997). Controlled acute and follow-up trial of cognitive therapy and pharmacotherapy in out-patients with recurrent depression. *British Journal of Psychiatry, 171,* 328–334.

Bower, P., Rowland, N., & Hardy, R. (2003). The clinical effectiveness of counselling in primary care: A systematic review and a meta-analysis. *Psychological Medicine, 33,* 203–215.

†Brown, R. A., & Lewinsohn, P. M. (1984). A psychoeducational approach to the treatment of depression: Comparison of group, individual, and minimal contact procedures. *Journal of Consulting and Clinical Psychology, 52,* 774–783.

Cohen, J. (1965). Some statistical issues in psychological research. In B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

Crits-Christoph, P. (1992). The efficacy of brief dynamic psychotherapy: A meta-analysis. *American Journal of Psychiatry, 149,* 151–158.

Derogatis, L. R. (1983). *SCL-90–R: Administration, scoring, and procedural manual—II.* Baltimore: Clinical Psychometric Research.

Derogatis, L. R., Rickels, K., & Rock, A. F. (1976). The SCL–90 and the MMPI: A step in the validation of a new self-report scale. *British Journal of Psychiatry, 128,* 280–289.

DeRubeis, R. J., Gelfand, L. A., Tang, T. Z., & Simons, A. D. (1999). Medications versus cognitive behavior therapy for severely depressed outpatients: Mega-analysis of four randomized comparisons. *American Journal of Psychiatry, 156,* 1007–1013.

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 57,* 414–419.

**Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., et al. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: I. General effectiveness of treatments. *Archives of General Psychiatry, 46,* 971–982.

Gaffan, E. A., Tsaousis, I., & Kemp-Wheeler, S. M. (1995). Researcher allegiance and meta-analysis: The case of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology, 63,* 966–980.

Garfield, S. L. (1986). Research on client variables in psychotherapy. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (3rd ed., pp. 213–256). New York: Wiley.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 339–355). New York: Russel Sage Foundation.

Gloaguen, V., Cottraux, J., Cucherat, M., & Blackburn, I. (1998). A meta-analysis of the effects of cognitive therapy in depressed patients. *Journal of Affective Disorders, 49,* 59–72.

Goldfried, M., & Wolfe, B. (1998). Toward a clinically valid approach to therapy research. *Journal of Consulting and Clinical Psychology, 66,* 143–150.

*Greenberg, L. S., & Watson, J. (1998). Experiential therapy of depression: Differential effects of client-centered relationship conditions and process experiential interventions. *Psychotherapy Research, 8,* 210–224.

Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology, 32,* 50–55.

Hamilton, M. A. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry, 23,* 56–62.

Hamilton, M. A. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology, 6,* 278–296.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* San Diego, CA: Academic Press.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods, 9,* 426–445.

**Hollon, S. D., DeRubeis, R. J., Evans, M. D., Wiemer, M. J., Garvey, M. J., Grove, W. M., & Tuason, V. B. (1992). Cognitive therapy and pharmacotherapy for depression: Singly and in combination. *Archives of General Psychiatry, 49,* 774–781.

Hollon, S. D., & Kendall, P. C. (1980). Cognitive self-statements in depression: Development of an Automatic Thoughts Questionnaire. *Cognitive Therapy and Research, 4,* 383–395.

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose–effect relationship in psychotherapy. *American Psychologist, 41,* 159–164.

*Jacobson, N. S., Dobson, K., Fruzzetti, A. E., Schmaling, K. B., & Salusky, S. (1991). Marital therapy as a treatment for depression. *Journal of Consulting and Clinical Psychology, 59,* 547–557.

**Jacobson, N. S., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., et al. (1996). A component analysis of cognitive–behavioral treatment for depression. *Journal of Consulting and Clinical Psychology, 64,* 295–304.

**Jarrett, R. B., Schaffer, M., McIntire, D., Witt-Browder, A., Kraft, D., & Risser, R. C. (1999). Treatment of atypical depression with cognitive therapy or phenelzine: A double-blind, placebo-controlled trial. *Archives of General Psychiatry, 56,* 431–437.

**Johnson, W. B., DeVries, R., Ridley, C. R., Pettorini, D., & Peterson, D. R. (1994). The comparative efficacy of Christian and secular rational–emotive therapy with Christian clients. *Journal of Psychology and Theology, 22,* 130–140.

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

Kim, D.-M., Wampold, B. E., & Bolt, D. M. (2006). Therapist effects in psychotherapy: A random-effects modeling of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Psychotherapy Research, 16,* 161–172.

**King, M., Sibbald, B., Ward, E., Bower, P., Lloyd, M., Gabbay, M., & Byford, S. (2000). Randomised controlled trial of non-directive counselling, cognitive–behaviour therapy and usual general practitioner care in the management of depression as well as mixed anxiety and depression in primary care. *Health Technology Assessment, 4,* 1–83.

*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 Pracitce, 9,* 213–229.

*Kornblith, S. J., Rehm, L. P., O'Hara, M. W., & Lamparski, D. M. (1983). The contribution of self-reinforcement training and behavioral assignments to the efficacy of self-control therapy for depression. *Cognitive Therapy and Research, 7,* 499–528.

Lambert, M. J. (2001). The status of empirically supported therapies: Comment on Westen and Morrison's (2001) multidimensional meta-analysis. *Journal of Consulting and Clinical Psychology, 69,* 910–913.

Lambert, M. J., & Bergin, A. E. (1994). The effectiveness of psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *The handbook of psychotherapy and behavior change* (pp. 143–189). New York: Wiley.

Lambert, M. J., Hatch, D. R., Kingston, M. D., & Edwards, B. C. (1986). Zung, Beck, and Hamilton Rating Scales as measures of treatment outcome: A meta-analytic comparison. *Journal of Consulting and Clinical Psychology, 54,* 54–59.

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed., pp. 139–193). New York: Wiley.

Luborsky, L. (1972). Research cannot yet influence clinical practice. In A. Bergin & H. Strupp (Eds.), *Changing frontiers in the science of psychotherapy* (pp. 120–127). Chicago: Aldine.

Luborsky, L., Rosenthal, R., Diguer, L., Andrusyna, T. P., Berman, J. S., Levitt, J. T., et al. (2002). The dodo bird verdict is alive and well—Mostly. *Clinical Psychology: Science and Practice, 9,* 2–12.

McCullough, M. E. (1999). Research on religion-accommodative counseling: Review and meta-analysis. *Journal of Counseling Psychology, 46,* 92–98.

*McKnight, D. L., Nelson-Gray, R. O., & Barnhill, J. (1992). Dexamethasone suppression test and response to cognitive therapy and antidepressant medication. *Behavior Therapy, 23,* 99–111.

*McNamara, K., & Horan, J. J. (1986). Experimental construct validity in the evaluation of cognitive and behavioral treatments for depression. *Journal of Counseling Psychology, 33,* 23–30.

Merrill, K. A., Tolbert, V. E., & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of Consulting and Clinical Psychology, 71,* 404–409.

Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (in press). Using clinical trials to benchmark effects produced in clinical practice. *Quality and Quantity.*

**Murphy, G. E., Simons, A. D., Wetzel, R. D., & Lustman, P. J. (1984). Cognitive therapy and pharmacotherapy: Singly and together in the treatment of depression. *Archives of General Psychiatry, 41,* 33–41.

*Neimeyer, R. A., & Feixas, G. (1990). The role of homework and skill acquisition in the outcome of group cognitive therapy for depression. *Behavior Therapy, 21,* 281–292.

*†Nezu, A. M. (1986). Efficacy of a social problem-solving therapy approach for unipolar depression. *Journal of Consulting and Clinical Psychology, 54,* 196–202.

*†Nezu, A. M., & Perri, M. G. (1989). Social problem-solving therapy for unipolar depression: An initial dismantling investigation. *Journal of Consulting and Clinical Psychology, 57,* 408–413.

*†Pecheur, D. R., & Edwards, K. J. (1984). A comparison of secular and religious versions of cognitive therapy with depressed Christian college students. *Journal of Psychology and Theology, 12,* 45–54.

Posternak, M. A., & Miller, I. (2001). Untreated short-term course of major depression: A meta-analysis of outcomes from studies using wait-list control groups. *Journal of Affective Disorders, 66,* 139–146.

*†Propst, L. R., Ostrom, R., Watkins, P., Dean, T., & Mashburn, D. (1992). Comparative efficacy of religious and nonreligious cognitive–behavioral therapy for the treatment of clinical depression in religious individuals. *Journal of Consulting and Clinical Psychology, 60,* 94–103.

Quintana, S. M., & Minami, T. (2006). Guidelines for meta-analyses of counseling psychology research. *Counseling Psychologist, 34,* 839–877.

*Rehm, L. P., Kaslow, N. J., & Rabin, A. S. (1987). Cognitive behavioral targets in a self-control therapy program for depression. *Journal of Consulting and Clinical Psychology, 55,* 60–67.

*†Rehm, L. P., Kornblith, S. J., O'Hara, M. W., Lamparski, D. M., Romano, J. M., & Volkin, J. I. (1981). An evaluation of major components in a self-control therapy program for depression. *Behavior Modification, 5,* 459–489.

Robinson, L. A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for treatment of depression: A comprehensive review of controlled outcome research. *Psychological Bulletin, 108,* 30–49.

Rogers, C. R. (1951). *Client-centered therapy: Its current practice, implications, and theory.* Boston: Houghton Mifflin.

†Rokke, P. D., Tomhave, J. A., & Jocic, Z. (2000). Self-management therapy and educational group therapy for depressed elders. *Cognitive Therapy and Research, 24,* 99–119.

*Roth, D., Bielski, R., Jones, M., Parker, W., & Osborn, G. (1982). A comparison of self-control therapy and combined self-control therapy and antidepressant medication in the treatment of depression. *Behavior Therapy, 13,* 133–144.

Rounsaville, B. J., O'Malley, S., Foley, S., & Weissman, M. M. (1988). Role of manual-guided training in the conduct and efficacy of interpersonal psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 56,* 681–688.

**Rush, A. J., Beck, A. T., Kovacs, M., & Hollon, S. (1977). Comparative efficacy of cognitive therapy and pharmacotherapy in the treatment of depressed outpatients. *Cognitive Therapy and Research, 1,* 17–37.

*Rush, A. J., & Watkins, J. T. (1981). Group versus individual cognitive therapy: A pilot study. *Cognitive Therapy and Research, 5,* 95–103.

*Scott, A. I., & Freeman, C. P. L. (1992). Edinburgh primary care depression study: Treatment outcome, patient satisfaction, and cost after 16 weeks. *British Medical Journal, 304,* 883–887.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The *Consumer Reports* Study. *American Psychologist, 50,* 965–974.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83.

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Shadish, W. R., Matt, G. E., Navarro, A. M., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin, 126,* 512–529.

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crits-Christoph, P., Hazeligg, M. D., et al. (1997). Evidence that therapy works in clinically representative conditions. *Journal of Consulting and Clinical Psychology, 65,* 355–365.

Shadish, W. R., Montgomery, L. M., Wilson, P., Wilson, M. R., Bright, I., & Okwumabua, T. (1993). Effects of family and marital psychotherapies: A meta-analysis. *Journal of Consulting and Clinical Psychology, 61,* 992–1002.

Shapiro, D. A., & Shapiro, D. (1982). Meta-analysis of comparative therapy outcome studies: A replication and refinement. *Psychological Bulletin, 92,* 581–604.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.

Spitzer, R. L., & Williams, J. B. W. (1984). *Structured Clinical Interview for DSM–III (SCID).* New York: New York State Psychiatric Institute Biometrics Research.

Spitzer, R. L., Williams, J. B. W., Gibbon, M., & First, M. B. (1989). *Structured Clinical Interview for DSM–III–R.* New York: New York State Psychiatric Institute Biometrics Research.

Spitzer, R. L., Williams, J. B. W., Gibbon, M., & First, M. B. (1995). *Structured Clinical Interview for DSM–IV.* New York: American Psychiatric Press.

**Steuer, J. L., Mintz, J., Hammen, C. L., Hill, M. A., Jarvik, L. F., McCarley, T., et al. (1984). Cognitive–behavioral and psychodynamic group psychotherapy in treatment of geriatric depression. *Journal of Consulting and Clinical Psychology, 52,* 180–189.

*Stravynski, A., Verreault, R., Gaudette, G., Langlois, T., Gagnier, S., & Larose, M. (1994). The treatment of depression with group behavioural–cognitive therapy and imipramine. *Canadian Journal of Psychiatry, 39,* 387–390.

Strupp, H. H. (1989). Psychotherapy: Can the practitioner learn from the researcher? *American Psychologist, 44,* 717–724.

Svartberg, M., & Stiles, T. C. (1991). Comparative effects of short-term psychodynamic psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology, 59,* 704–714.

Thase, M. E., Greenhouse, J. B., Frank, E., Reynolds, C. F., III, Pilkonis, P. A., Hurley, K., et al. (1997). Treatment of major depression with psychotherapy or psychotherapy–pharmacotherapy combinations. *Archives of General Psychiatry, 54,* 1009–1015.

**Thompson, L. W., Coon, D. W., Gallagher-Thompson, D., Sommer, B. R., & Koin, D. (2001). Comparison of desipramine and cognitive/behavioral therapy in the treatment of elderly outpatients with mild-to-moderate depression. *American Journal of Geriatric Psychiatry, 9,* 225–240.

*Thompson, L. W., & Gallagher, D. (1984). Efficacy of psychotherapy in the treatment of late-life depression. *Advances in Behaviour Research and Therapy, 6,* 127–139.

† Thompson, L. W., Gallagher, D., & Breckenridge, J. S. (1987). Comparative effectiveness of psychotherapies for depressed elders. *Journal of Consulting and Clinical Psychology, 55,* 385–390.

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

Wade, W. A., Treat, T. A., & Stuart, G. L. (1998). Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy. *Journal of Consulting and Clinical Psychology, 66,* 231–239.

Wampold, B. E. (2001). *The great psychotherapy debate: Model, methods, and findings.* Mahwah, NJ: Erlbaum.

Wampold, B. E., Minami, T., Baskin, T. W., & Tierney, S. C. (2002). A meta-(re)analysis of the effects of cognitive therapy versus "other therapies" for depression. *Journal of Affective Disorders, 68,* 159–165.

Wampold, B. E., Mondin, G. W., Moody, M., Stich, F., Benson, K., & Ahn, H. (1997). A meta-analysis of outcome studies comparing bona fide psychotherapies: Empirically, "All must have prizes." *Psychological Bulletin, 122,* 203–215.

**Watson, J. C., Gordon, L. B., Stermac, L., Kalogerakos, F., & Steckley, P. (2003). Comparing the effectiveness of process-experiential with cognitive–behavioral psychotherapy in the treatment of depression. *Journal of Consulting and Clinical Psychology, 71,* 773–781.

Weersing, V. R., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology, 70,* 299–310.

Westen, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69,* 875–899.

†Wilson, P. H., Goldin, J. C., & Charbonneau-Powis, M. (1983). Comparative efficacy of behavioral and cognitive treatments of depression. *Cognitive Therapy and Research, 7,* 111–124.

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

Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12,* 63–70.

Received April 11, 2005
Revision received December 18, 2006
Accepted December 27, 2006 ■