Psychotherapy for the Treatment of Depression: A Comprehensive Review of Controlled Outcome Research

Leslie A. Robinson, Jeffrey S. Berman, and Robert A. Neimeyer
Memphis State University

Previous quantitative reviews of research on the efficacy of psychotherapy for depression have included only a subset of the available research or limited their focus to a single outcome measure. The present review offers a more comprehensive quantitative integration of this literature. Using studies that compared psychotherapy with either no treatment or another form of treatment, this article assesses (a) the overall effectiveness of psychotherapy for depressed clients, (b) its effectiveness relative to pharmacotherapy, and (c) the clinical significance of treatment outcomes. Findings from the review confirm that depressed clients benefit substantially from psychotherapy, and these gains appear comparable to those observed with pharmacotherapy. Initial analysis suggested some differences in the efficacy of various types of treatment; however, once the influence of investigator allegiance was removed, there remained no evidence for the relative superiority of any 1 approach. In view of these results, the focus of future research should be less on differentiating among psychotherapies for depression than on identifying the factors responsible for improvement.

Depression is a prevalent clinical disorder with high economic and emotional costs. Epidemiological research has indicated that 10%–20% of the population experience a major depressive episode at some point in their lifetime (Boyd & Weissman, 1981), with the incidence highest during the adult years when family and career responsibilities may be most adversely affected (Weissman & Myers, 1978). Although the remission rate for depressive disorders is relatively high (Beck, 1967, chap. 3), a substantial portion of those afflicted remain chronically depressed (Weissman & Klerman, 1977), and those who do improve are at an increased risk for further episodes (Belsher & Costello, 1988; Kessler, 1978; Klerman, 1978).

Until recently, depression was treated almost exclusively with medication, traditional insight-oriented therapy, or a combination of the two. However, the 1970s witnessed the development of a number of new therapeutic approaches, each of which possessed a different etiological model for depression. For example, advocates of behavioral approaches treat depression as a consequence of a low rate of response-contingent positive reinforcement. The object of this therapy, then, is to increase reinforcement, either by encouraging participation in pleasant activities (Lewinsohn, 1974) or by building the assertion skills necessary to elicit social rewards (LaPointe & Rimm, 1980; Sanchez, Lewinsohn, & Larson, 1980). A second treatment approach, cognitive therapy, is derived from Beck’s (1967) view of depression as an affective response to negative beliefs. Modifying these unproductive beliefs is the primary focus of cognitive therapy. In addition, a number of treatment packages have been developed that explicitly integrate elements from both cognitive and behavioral models. Examples include Lewinsohn’s Coping with Depression course (Lewinsohn, Steinmetz, Antonuccio, & Teri, 1985) and Rehm’s self-control therapy (Fuchs & Rehm, 1977).

With the development of these new therapies has come a dramatic increase in outcome research on the efficacy of treatments for depression (Brown & Lewinsohn, 1984; Elkin, Parloff, Hadley, & Autry, 1985; McLean, 1981). Two major evaluation strategies have been used: (a) comparisons between treated clients and wait-list controls and (b) comparisons between clients receiving different types of psychotherapy. The results of studies using wait-list comparisons have generally indicated that a number of treatment approaches are effective. However, as Kazdin (1981) has pointed out, the focus in recent depression research has been not so much on comparisons with untreated controls as on comparisons between different types of therapy, and here the results have been remarkably inconsistent. For example, Shaw (1977) found that cognitive therapy was more effective than behavioral treatment, whereas others (e.g., Hodgson, 1981; LaPointe & Rimm, 1980) have reached the opposite conclusion. Similarly mixed results have been obtained when cognitive or behavioral methods have been compared...
with more traditional approaches (e.g., Fleming & Thornton, 1980; LaPointe & Rimm, 1980; Thompson & Gallagher, 1984). To further complicate matters, few studies have obtained consistent results across all outcome measures or assessment points.

In attempting to account for these inconsistencies, researchers have advanced two primary hypotheses. First, it has been suggested that genuine differences in efficacy do exist but have been obscured by variations across studies in factors such as treatment procedures, client selection, and therapist training. According to this view, consistent differences in therapy outcome might emerge in large-scale, tightly controlled research programs. Such is the philosophy underlying the Treatment of Depression Collaborative Research Program (Elkin et al., 1985), a multisite project initiated by the National Institute of Mental Health to evaluate the efficacy of cognitive–behavioral, interpersonal, and pharmacological treatments for depression.

In contrast, others have suggested that there are no significant differences in the effects of the various therapies, partly because of considerable overlap in their treatment methods. Both Kazdin (1981) and McLean (1982) have pointed out that there may well be no treatment techniques that are used only within a particular therapy. Instead,

what researchers and clinicians presently have . . . is broad agreement on the characteristics of clinical depression . . . ; divergent theory to account for the hypothesized mechanisms responsible for the etiology, maintenance, and reversal of depression; and strikingly similar treatment procedures deriving from these diverse theories. (McLean, 1982, p. 22)

If the methods and effects of different forms of psychotherapy are similar, why then do some studies reveal reliable differences in treatment efficacy, whereas others do not? One possible explanation involves the impact of the researcher’s allegiance. In previous reviews of psychotherapy research (Berman, Miller, & Massman, 1985; Smith, Glass, & Miller, 1980, chap. 5), the results of comparisons between therapies have been found to vary according to the theoretical preference of the investigator. Although it seems reasonable to expect that a similar effect might be operative in the depression literature, this possibility has not yet been explored.

The typical response to the inconsistency in the existing literature has been a call for further research, in the hope that better designed studies might reveal some client or treatment format variables that could account for the contradictory findings of previous research. Additional studies would undoubtedly contribute to the literature, but it appears unlikely that they can provide any final answers to the central problems in depression research. As Fiske (1983) has pointed out, although “the single study may stimulate or irritate in a healthy fashion, only the distillations from the entire body of research in an area have lasting effects” (p. 65).

This type of broad integration of the depression treatment literature has been attempted in a number of recent narrative reviews. However, as of yet no clear and consistent conclusions have been reached. For example, some commentators have indicated that the research supports the efficacy of both cognitive and behavioral approaches to the treatment of depression (e.g., Blaney, 1981; Emmelkamp, 1986; McLean, 1982; Rehm & Kornblith, 1979; Whitehead, 1979), whereas others have concluded that behavioral techniques have not yet received adequate empirical support (e.g., DeRubeis & Hollon, 1981; Hollon, 1981; Kovacs, 1980). In a few instances, it has been suggested that treatments incorporating both cognitive and behavioral components may be more effective than either approach alone (e.g., Blaney, 1981; Rehm & Kornblith, 1979; Whitehead, 1979). Other reviewers have simply concluded that at this point, there is no clear treatment of choice (e.g., Emmelkamp, 1986; McLean, 1982; Rush, 1982). Furthermore, although differences in client populations and therapy formats have often been cited as contributing to inconsistent results across studies, few reviewers have systematically examined research relating these variables to outcome. Most often, they have simply noted that there are at present no clear prognostic or prescriptive indicators and have called for further research in these areas (e.g., Blaney, 1981; DeRubeis & Hollon, 1981; Kovacs, 1980; McLean, 1981; Rush, 1982; Whitehead, 1979).

Given the large number of studies on the treatment of depression and the complexity of their results, other reviewers have turned to quantitative techniques for summarizing the research. The advantage of these quantitative, or “meta-analytic,” procedures (e.g., see Glass, McGaw, & Smith, 1981) is that they provide a powerful method for identifying trends that might otherwise be overlooked. In addition, these techniques permit the systematic assessment of factors that vary across individual studies.

Although several previous quantitative reviews of the psychotherapy literature (Dush, Hirt, & Schroeder, 1983; Shapiro & Shapiro, 1982; Smith et al., 1980) have reported separate analyses of depressed samples, in each case these analyses were limited to a small sample of the research on psychotherapy treatments for depression. Three other reviews that focused specifically on depression have suffered from a similar limitation. In two of these reviews (Quality Assurance Project, 1983; Steinbrueck, Maxwell, & Howard, 1983), the primary purpose was to estimate the relative benefits of pharmacotherapy and psychotherapy for depression; the third review (Conte, Plutchik, Wild, & Karasu, 1986) used a “box-score” summary technique to evaluate treatments combining pharmacotherapy with psychotherapy. Given their focus on drug treatments, however, these reviews included relatively few studies that assessed the efficacy of psychotherapy alone. Thus, many substantive issues concerning psychotherapy for depression could not be addressed.

In contrast, two recent reviews (Dobson, 1989; Nietzel, Russell, Hemmings, & Grettet, 1987) have included substantially larger portions of the literature on psychotherapy for depression. However, even these reviews were less than comprehensive, because they were both limited to studies that used the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) as an outcome measure, and only the BDI was analyzed. Moreover, the two reviews reached markedly different conclusions. Nietzel et al. (1987) found that individual treatment was more effective than group treatment, but there were no reliable differences in the efficacy of cognitive, behavioral, and other forms of therapy. In contrast, Dobson (1989) concluded that Beck’s cognitive therapy was more effective than other therapeutic approaches. Perhaps one reason for these di-
vergent findings is that neither review adjusted for the potential influence of investigator allegiance in the studies they examined.

The purpose of this review was to provide a more complete quantitative summary of the controlled research evidence on psychotherapy for depression. In the first of the following analyses, we examined the efficacy of psychotherapy for depression by using evidence from a substantially larger number of empirical studies than have been included in any previous quantitative review of this literature. In a second set of analyses, we used the available research to assess the comparative efficacy of psychotherapy and the leading alternative treatment for depression, pharmacotherapy. Finally, we evaluated the clinical significance of psychotherapy by comparing treated clients with nondepressed individuals. Thus, our analyses allowed us to assess the effectiveness of psychotherapy for depression, its benefits relative to pharmacotherapy, and how much former clients resemble those who are not depressed.

Analysis 1: Effectiveness of Psychotherapy

In this first set of analyses, we investigated not only the overall effectiveness of psychotherapy but also the relative effectiveness of different forms of treatment. In addition, we examined a number of other substantive issues in research on the treatment of depression, including the role of investigator allegiance, the impact of group and individual therapy formats, the importance of diagnostic screening procedures, and the influence of other variables such as therapist training, length of treatment, and client characteristics.

Method

Studies. The first analysis was based on a total of 58 studies of psychotherapy for the treatment of depression. (See Appendix A for a list of the references for these studies.) The studies were identified through a search of the volumes of Psychological Abstracts (1976–1986), references of published reviews and outcome studies, and an issue-by-issue examination of the 1985 and 1986 volumes of relevant journals. Of the 58 studies included in this analysis, 47 were not included in the recent Dobson (1989) review and 40 were not covered in the analyses by Nietzel et al. (1987).

We used several criteria in selecting studies for inclusion in the review. First, to assess the effects of therapy on depressive disorders (rather than depressive moods), the analysis was restricted to studies using samples identified as primarily suffering from depression. Thus, studies that described clients in more general terms (e.g., neurotic) or in terms of another specific diagnostic category (e.g., alcoholic) were excluded, even when the researcher reported that the clients were also depressed. Studies using inpatient samples and those that focused on children or adolescents were also omitted, because treatment methods used with these groups often differ from those that are the focus of this review.

Second, we included a study in the review only if it contained a comparison between treatment and no treatment or between different types or formats of therapy. Thus, we omitted case histories and studies using simple pre-post designs.

Third, because our primary interest was in the effects of psychotherapy, we excluded research on treatments that did not have a prominent verbal component. Thus, treatments such as exercise and bibliotherapy were not considered. In addition, we omitted the very few studies of family and marital therapy because their interactional focus differed substantially from the more individualized goals of the treatments included in the reviewed studies.

The studies selected for review investigated a variety of psychotherapeutic methods. Distinctions between the different types of therapy were often difficult to draw, in part because of overlap in the treatment techniques. However, a careful inspection of the method sections of these studies indicated that most therapies could be classified into one of four categories: (a) cognitive, (b) behavioral, (c) cognitive–behavioral, and (d) general verbal therapy. The first category, cognitive therapy, included those treatments that focused primarily on the evaluation and modification of cognitive patterns. For example, treatments that involved attributional retraining or challenging irrational beliefs were classified as cognitive. However, therapies that simply directed clients to substitute positive thoughts or images for negative ones were excluded from this category. Such treatments differ from cognitive therapy as it is usually practiced in that no evaluation of existing cognitions is undertaken by the client (Ledwidge, 1978; also see Miller & Berman, 1983). The behavioral therapy category included treatments designed to decrease depression by changing behavioral patterns (e.g., by increasing assertive behavior or participation in pleasant activities). Therapies that included both cognitive and behavioral components were classified as cognitive–behavioral treatments. The final category, general verbal therapy, comprised treatments such as psychodynamic therapy, client-centered approaches, and other forms of interpersonal therapy such as that outlined by Klerman, Weissman, Rounsaville, and Chevron (1984). The commonality among these treatments is that each places relatively greater emphasis on insight than on the acquisition of a set of specific skills. Although the general verbal category was broad, too few studies were available to evaluate the effectiveness of the specific therapies within this group.

Treatment outcomes were assessed in the studies by a variety of instruments. Many of the instruments were designed specifically to assess depressive symptomatology, but some were more general or evaluated other areas of functioning. We classified the following scales as specific measures of depression: the BDI (Beck et al., 1961), the Zung Self-Rating Depression Scale (Zung, 1965), the Depression Adjective Check List (Lubin, 1965), the Hamilton Rating Scale for Depression (Hamilton, 1960), the Depression Scale of the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1967), the D-30 (Dempsey, 1964), and the Center for Epidemiologic Studies Depression Scale (Radloff, 1977).

Descriptive characteristics of the 58 studies reviewed are presented in Table 1. As the table reveals, the typical client was a middle-aged woman who was experiencing moderate depression as measured by the BDI. The therapy was usually brief, and treatment sessions occurred approximately once a week. In the 34 studies that included a follow-up assessment, this follow-up was conducted an average of 13 weeks (range = 2–52) after treatment termination.

In 28 (48%) of the studies, clients were solicited from the community through media announcements. Another 14 investigations (24%) relied on students solicited in a university setting, and 9 studies (16%) used traditionally referred outpatients. In the remaining 7 studies (12%), either the referral source was not reported or both solicited and traditionally referred clients were included.

Although all studies focused on the treatment of depression, some screened clients more rigorously than did others. In 20 investigations (35%), clients were required to meet formal diagnostic criteria for a depressive disorder in order to be included. Most often, the Research

Computing such effect sizes are reported in previous work (e.g., Miller & Berman, 1983; also see Glass et al., 1981, chap. 5). Occasionally, researchers described a measure but did not report findings for it or reported results only as nonsignificant. In these situations, an exact effect size could not be calculated. However, excluding these measures would have inflated the estimate of overall effect size, because investigators are more likely to provide adequate information on measures that reveal reliable group differences. Thus, when findings were not reported or were described simply as nonsignificant, we conservatively estimated the effect size to be zero.

Hedges (1982) has identified a small-sample bias in the estimate of effect size. Although this bias is of practical concern only when sample sizes are quite small (i.e., fewer than 20 subjects), we applied Hedges' (1982, Formula 4) correction for the bias to all effect sizes reported in our analyses.

Preliminary analyses. Differences between groups were usually assessed by more than one outcome measure (M = 6.0, range = 1–25). Moreover, many studies reported results for more than one treatment comparison (M = 2.1, range = 1–6). Other reviewers have treated the effect sizes derived from individual outcome measures as separate observations in their analyses (e.g., see Smith et al., 1980, chap. 4). However, this procedure arbitrarily weights studies according to the number of outcome measures and treatment comparisons reported. Even worse, multiple effect sizes derived from the same study may not represent statistically independent observations. If effect sizes within a study are not independent, then using the effect size as the unit of analysis can seriously underestimate error variance and inflate tests of statistical significance (e.g., see Glass et al., 1981, chap. 6).

To assess the issue of nonindependence in our data, we first examined whether the variation in effect sizes derived from different treatment comparisons was greater than the variation of effect sizes within a single treatment comparison. Our focus in this analysis was on the most frequent type of comparison, that between treated clients and untreated controls. We conducted an analysis of variance in which the unit of analysis was the individual effect size (N = 354) and the independent variable was the treatment comparison (N = 78). This analysis revealed that the variability among effect sizes drawn from different treatment comparisons was indeed far greater than the variability of effect sizes within a treatment comparison, intraclass $R = .66, F(77, 276) = 9.68, p < .001$.

Given this problem of nonindependence, we averaged all effect sizes derived from the same treatment comparison. For example, if a treatment comparison yielded five different effect sizes, the mean of these measures was calculated for use in overall analyses. However, separate means for different types of outcome measures were also retained. For instance, we calculated the mean of effect sizes derived from self-report outcome measures and the mean of those based on observer ratings. These additional means could then be used in analyses in which we wished to assess treatment effects for particular types of outcome measures.

Many of the studies reported the results of more than one treatment comparison. As a further check on nonindependence, we therefore examined whether the variability among treatment comparisons drawn from different studies was greater than the variability of multiple treatment comparisons within the same study. As before, our focus was on the most frequent type of comparison, that between treated clients and untreated controls. We conducted an analysis of variance in which the unit of analysis was the mean effect size for a comparison (N = 78) and the independent variable was the study (N = 37). This analysis indicated that the variability among comparisons from different studies was sub-

---

Table 1

<table>
<thead>
<tr>
<th>Characteristics of Psychotherapy Studies</th>
<th>M</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of clients</td>
<td>40.4</td>
<td>9–155</td>
</tr>
<tr>
<td>Clients per group</td>
<td>14.3</td>
<td>4–47</td>
</tr>
<tr>
<td>Percentage attrition (posttreatment)</td>
<td>11.0</td>
<td>0–65</td>
</tr>
<tr>
<td>Percentage female clients</td>
<td>79.6</td>
<td>50–100</td>
</tr>
<tr>
<td>Client age (years)</td>
<td>39.4</td>
<td>19–71</td>
</tr>
<tr>
<td>Initial Beck Depression Inventory score</td>
<td>22.7</td>
<td>12–30</td>
</tr>
<tr>
<td>No. of therapists</td>
<td>4.0</td>
<td>1–18</td>
</tr>
<tr>
<td>Weeks of treatment</td>
<td>6.9</td>
<td>1–36</td>
</tr>
<tr>
<td>No. of sessions</td>
<td>8.7</td>
<td>1–46</td>
</tr>
</tbody>
</table>

Note. Each mean is based on at least 42 of the 58 studies.

Diagnostic Criteria (Spitzer, Endicott, & Robins, 1978) were used, but other investigators used the Feighner criteria (Feighner et al., 1972) or the Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 1980). The 38 other studies (65%) used less stringent selection criteria, such as scores on self-report measures of depression.

**Estimating treatment effects.** Each outcome measure reported in a study was expressed in terms of Cohen's (1977) $d$, a standardized measure of effect size. Cohen's $d$ is defined as

$$d = \frac{m_1 - m_2}{s},$$

where $m_1$ and $m_2$ represent group means and $s$ is the pooled within-groups standard deviation. For comparisons between controls and treated clients, the control mean was subtracted from the treatment mean. Thus, an effect size of 0.5 indicated that the mean of the treated group was one half of a standard deviation larger than the mean of the control group. In direct comparisons between different types or formats of treatment, the difference in means was calculated in a consistent manner. For example, we always calculated effect sizes for comparisons between cognitive and behavioral therapy by subtracting the behavioral group mean from the cognitive mean. For some outcome measures, higher numbers indicated greater improvement, whereas for other measures lower numbers represented greater improvement. The sign of each effect size was therefore adjusted so that positive values indicated that clients in the first group improved more and negative values indicated that clients in the second group improved more.

We calculated effect sizes for all of the measures within a study, with a few exceptions. Process measures, such as client ratings of therapist likability, were not coded, because these measures are not a direct indication of treatment efficacy. Measures of paralinguistic characteristics (e.g., the latency of client speech) and observations of body movement (e.g., counts of head movements) were also excluded, because as Jacobson (1981) has pointed out, the validity of these data in the assessment of outcome has not been demonstrated. For similar reasons, the few measures of physiological characteristics, such as heart rate, were omitted. Measures administered midway through treatment were not coded. Also, when multiple follow-up assessments were conducted, only measures taken at the longest follow-up point were used to compute follow-up effect sizes.

Most effect sizes were calculated from means and standard deviations reported in the studies. When this information was not available, we used other statistics to estimate the effect size. The various methods for computing such effect sizes are reported in previous work (e.g., Miller & Berman, 1983; also see Glass et al., 1981, chap. 5).
substantially larger than the variation of comparisons within the same study, intraclass \( R = .76, F(36, 41) = 7.67, p < .001 \).

Because of this second level of nonindependence, we combined the results of multiple treatment comparisons from the same study. For example, if a study yielded three different comparisons of treatment to no treatment, we averaged the results of these comparisons for our overall analyses. However, we also retained separate means for specific types of treatment. For example, we calculated the mean effect size within a study for all comparisons involving cognitive therapy or all comparisons involving behavioral therapy. These additional means could then be used in analyses comparing the effectiveness of particular types or forms of treatment.

The goal of averaging multiple effect sizes within a treatment comparison and of averaging multiple comparisons within a study was to ensure that estimates of error variance were always based on independent observations. In most cases, this procedure meant that no matter how many effect sizes were derived from a study, the study provided only one observation in each of our statistical analyses. However, there were two types of analysis in which a study was represented by more than one score. The first and simpler case occurred when a group of studies provided an effect size for each of two categories being compared. For example, some of the studies provided an effect size at both posttreatment and follow-up, and we wanted to use these studies to assess whether posttreatment and follow-up effects differed. In such cases, the dependency among pairs of scores was accommodated in a straightforward fashion by means of a paired \( t \) test.

A more complex situation arose, however, when some studies in an analysis provided an effect size for only one of the categories being compared, but other studies offered effect sizes for more than one of the categories. For instance, most studies examined either an individual treatment or a group treatment, but a small number of studies examined both treatment formats. In the overall analysis comparing the effectiveness of individual and group therapies, we permitted studies with information about both treatment formats to provide one observation to each treatment category. Thus, a study might contribute more than one score to the analysis but would provide no more than one observation to each treatment category. In this type of analysis, the estimate of variability within groups would still be based on independent observations. However, given the clustering of effect sizes drawn from the same study, the estimate of variability between groups would be smaller than that expected if all the observations were from separate studies (see Kenny & Judd, 1986, Equation 1). Therefore, these hybrid analyses represent conservative inferential tests.

The studies under review varied widely in terms of the number of clients in their samples. Analysis revealed that studies with fewer clients yielded larger overall treatment effects than studies with more clients, \( r(35) = -.32, p = .06 \). The most likely explanation for this pattern is that it reflects a bias in the publication of research findings. That is, journal editors may be more willing to publish studies that report statistically significant differences, and in studies with small samples only large treatment effects are statistically significant.

One might also expect that the variability in the findings of studies would decrease as the sample sizes of the studies increased. To evaluate this issue, we conducted a regression analysis in which we used the sample size of the study to predict the overall treatment effect. As expected, the squared residuals or errors from this regression analysis decreased as the sample size of the study increased, \( r(35) = -.35, p = .03 \). Therefore, studies with larger samples not only provided more conservative estimates of treatment effects but also offered more reliable estimates. Given this difference in the reliability of studies with different sample sizes, we conducted the following analyses using a weighted least squares procedure (e.g., see Chatterjee & Price, 1977, chap. 5) in which the results of each study were weighted by the number of clients in that study. Thus, the means, standard deviations, and correlations reported in the following sections are all weighted estimates.

**Results**

Table 2 presents the effect sizes for comparisons between psychotherapy and no treatment. As can be seen, posttreatment outcomes of depressed clients receiving psychotherapy were almost three fourths of a standard deviation better than the outcomes of individuals not receiving therapy. These posttreatment results were quite similar to those obtained with follow-up measures of outcome, \( t(44) = 0.20, p = .8 \). In fact, among the nine studies reporting both posttreatment and follow-up data, effect sizes derived from posttreatment measures \( (M = 0.74, SD = 0.61) \) were almost identical to effect sizes derived from the same measures at follow-up \( (M = 0.70, SD = 0.67) \), \( t(8) = 0.79, p = .5 \). Moreover, the posttreatment effect size for a study was highly predictive of the effect size for that study at follow-up, \( r(7) = .98, p < .001 \). Given this similarity between posttreatment and follow-up findings, the results presented in the remainder of the review are based only on posttreatment data.

The no-treatment control groups in the preceding analyses included not only wait-list controls but also placebo treatments, which provided attention or used placebo drugs. Analysis indicated that effect sizes based on comparisons to wait-list controls were larger than those based on comparisons to placebo controls, \( t(36) = 2.24, p = .03 \). As can be seen in Table 3, effect sizes for comparisons to wait-list controls differed reliably from zero, but effect sizes for comparisons to placebo controls did not. In view of these findings, we chose not to combine the two types of control groups in the remaining analyses. Instead, effect sizes involving comparisons to untreated individuals are based on the more frequently occurring type of control, the wait-list group.

**Table 3**

<table>
<thead>
<tr>
<th>Control group</th>
<th>( N ) of studies</th>
<th>( M )</th>
<th>( SD )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wait list</td>
<td>29</td>
<td>0.84*</td>
<td>0.69</td>
</tr>
<tr>
<td>Placebo</td>
<td>9</td>
<td>0.28</td>
<td>0.52</td>
</tr>
</tbody>
</table>

*Note.* Means and standard deviations are based on weighted least squares analyses in which effect sizes were weighted by sample size. * \( p < .05 \).
Type of treatment. Table 4 presents the effects of four specific types of psychotherapy: (a) cognitive, (b) behavioral, (c) cognitive-behavioral, and (d) general verbal. For each type of therapy, treated clients improved more than did wait-list controls. The effect sizes for some therapies were larger than for others, but this variation was not reliable, $F(3, 35) = 0.92, p = .4$. However, it is difficult to judge the relative efficacy of the treatments from this evidence, because the results were drawn from different sets of studies. For instance, studies evaluating general verbal therapy might have used clients that were less amenable to treatment than the client samples used in studies of other types of therapy. Thus, differences in effect size may be due to variations across studies in background variables such as sample characteristics rather than to differences in therapeutic efficacy.

Investigations that directly compare two or more types of treatment provide a better assessment of relative efficacy, because background variables are held relatively constant across the therapies. Thus, any differences between the treatment groups can be attributed more clearly to differences in the therapies themselves. Table 5 presents the results of studies that directly compared two or more therapies. Effect sizes were coded so that positive numbers indicate that the first therapy in the comparison was more effective, whereas negative numbers indicate that the second therapy was more effective. As the effect sizes in the table reveal, outcomes for cognitive therapy did not appear to differ reliably from those of either a strictly behavioral approach or a cognitive-behavioral intervention. However, cognitive-behavioral treatments did appear to produce more improvement than behavioral methods used in isolation. Furthermore, general verbal therapy seemed less effective than all three of the alternative therapies with which it was compared.

Although these results suggest that some forms of therapy may be more successful in alleviating depression than others, they do not take into account the possible impact of the researcher's allegiance. Previous research has indicated that when one therapy is compared with another, the preference of the researcher can have a sizable impact on a study's outcome, the opposite allegiance could often be rated on the basis of subtler cues. For example, one type of treatment might be described in considerable detail, whereas the other was only briefly and incompletely discussed. Alternatively, the introductory comments may have presented a balanced picture of the two therapies under comparison. This evenhanded treatment often occurred when researchers of differing theoretical persuasions collaborated. In such cases, the researchers were rated as having no preference.

Examination of the allegiance ratings suggested that investigators frequently had substantial theoretical preferences for particular forms of therapy. Furthermore, these ratings of allegiance were highly correlated with the results of direct comparisons between treatments, $r(28) = .58, p < .001$. Thus, therapies that were preferred by the investigator tended to achieve better results than the less favored therapies with which they were compared.

Although this finding might indicate that researcher allegiance had a sizable impact on a study's outcome, the opposite could have also been true. That is, study outcome could have influenced ratings of allegiance, because these ratings were based on introductory comments that were often prepared after a study was completed. To clarify the nature of the relation between allegiance and outcome, we conducted a second analysis that was restricted to comparisons in which the researchers' allegiance could be verified on the basis of their comments in some previous publication. In this smaller set of 13 studies, preferred therapies once again tended to produce more improvement than their less favored counterparts, $r(11) = .51, p = .08$.

### Table 4
**Efficacy of Different Types of Psychotherapy**

<table>
<thead>
<tr>
<th>Therapy</th>
<th>N of studies</th>
<th>Effect size M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive</td>
<td>8</td>
<td>0.96*</td>
<td>0.64</td>
</tr>
<tr>
<td>Behavioral</td>
<td>12</td>
<td>0.02*</td>
<td>0.79</td>
</tr>
<tr>
<td>Cognitive-behavioral</td>
<td>13</td>
<td>0.85*</td>
<td>0.74</td>
</tr>
<tr>
<td>General verbal</td>
<td>6</td>
<td>0.49*</td>
<td>0.35</td>
</tr>
</tbody>
</table>

Note. Means and standard deviations are based on weighted least squares analyses in which effect sizes were weighted by sample size. $* p < .05$.

### Table 5
**Direct Comparisons Between Different Types of Psychotherapy**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>N of studies</th>
<th>Effect size</th>
<th>Estimate if no allegiance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive vs. behavioral</td>
<td>12</td>
<td>0.12</td>
<td>0.12 (0.09)</td>
</tr>
<tr>
<td>Cognitive vs. cognitive-behavioral</td>
<td>4</td>
<td>-0.03</td>
<td>-0.03 (0.12)</td>
</tr>
<tr>
<td>Behavioral vs. cognitive-behavioral</td>
<td>8</td>
<td>-0.24*</td>
<td>-0.16 (0.10)</td>
</tr>
<tr>
<td>Cognitive vs. general verbal</td>
<td>7</td>
<td>0.47*</td>
<td>0.30</td>
</tr>
<tr>
<td>Behavioral vs. general verbal</td>
<td>14</td>
<td>0.27*</td>
<td>0.33</td>
</tr>
<tr>
<td>Cognitive-behavioral vs. general verbal</td>
<td>8</td>
<td>0.37*</td>
<td>0.38</td>
</tr>
</tbody>
</table>

Note. Means, standard deviations, and standard errors are based on weighted least squares analyses in which effect sizes were weighted by sample size.

* Positive numbers indicate that the first therapy in the comparison was more effective; negative numbers indicate that the second therapy in the comparison was more effective.

* Standard error of the estimated effect size is in parentheses.

* $p < .05$.

3 The two raters were able to judge allegiance on the basis of the introductory comments with a high degree of reliability. In fact, the intraclass correlation for the reliability of the mean of the two raters was .95.
Therefore, it seems unlikely that the relation between outcome and allegiance is solely attributable to the tendency of researchers to write introductions tailored to their results.

What would be the relative efficacy of the different therapies in the absence of an allegiance effect? To answer this question, we used the regression equation for the relation between effect size and allegiance rating to estimate the effect size that would have occurred if the investigators had no preference for either of the therapies being compared. The estimated mean effect size and its standard error are presented for each comparison in the final two columns of Table 5. It should be noted that the effect sizes for comparisons between cognitive and behavioral therapy and for those between cognitive and cognitive–behavioral approaches required no correction, because for these comparisons the investigators indicated no preference. However, for the four remaining comparisons in which there were substantial theoretical preferences, the effect sizes representing differences in treatment efficacy were substantially reduced. In fact, once the influence of investigator allegiance was removed, there remained no evidence for the relative superiority of any one type of therapy.

Some might wonder whether the lack of differences in treatment outcome is simply a reflection of incorrect treatment implementation. When studies do not verify that competing therapies were administered as prescribed, similarity in outcome may indicate either that the therapies were equally effective or that they were not implemented properly. In recognition of this issue, a number of investigators have used videotapes, audiotapes, or observers to ensure that the therapies were delivered in an appropriate manner. If the similarity in outcome across therapies is simply a reflection of poor treatment implementation, then one might expect the use of these monitoring procedures to increase the differences observed between treatments. However, in direct comparisons between treatments, our analyses revealed no reliable difference in the absolute magnitude of effect sizes from 14 studies that used these monitoring procedures (M = 0.27, SD = 0.15) and 16 studies that did not (M = 0.38, SD = 0.41), t(28) = 1.04, p = .3.

Another method of attempting to standardize treatment delivery involves the use of therapy manuals. In some of the studies, therapists conducted treatment according to detailed manuals; in other studies, no such formalized therapy program was used. When treatments were directly compared with one another, however, the absolute magnitude of effect sizes from 11 studies that used formal manuals (M = 0.28, SD = 0.30) did not differ reliably from the absolute magnitude of effect sizes from 14 studies in which no manuals were used (M = 0.34, SD = 0.18), t(23) = 0.55, p = .6. Similar results were observed when we examined studies comparing treated groups with wait-list controls. The effect sizes of 14 studies using manual-driven therapies (M = 0.82, SD = 0.64) did not differ systematically from the effect sizes of 17 studies for which no manual was developed (M = 0.84, SD = 0.74), t(29) = 0.07, p = .9. Although the use of treatment manuals has increased in recent years, these data provide no indication that their use either increases therapeutic efficacy or allows for a finer differentiation of the relative effectiveness of treatments.

The length of treatment might also affect the outcome of psychotherapy. To address this issue, we correlated the results of comparisons between treated groups and wait-list controls with the length of treatment. The analyses revealed no reliable relation between effect size and either the number of weeks of treatment, r(26) = −.09, p = .6, or the total number of sessions, r(27) = −.11, p = .6. Thus, there is no indication from the studies that the benefits of psychotherapy increase systematically as the length of treatment increases.

**Treatment format.** In some investigations clients were treated individually, whereas in others clients were seen in groups. To determine the possible influence of these different formats on outcome, we analyzed studies comparing individual therapy with a wait-list control and studies comparing group therapy with a wait-list control separately. As indicated in Table 6, effect sizes for the group and individual methods were quite similar, r(29) = 0.03, p > .9. Further data relevant to this point were provided by five investigations that directly compared individual and group approaches. From this smaller set of studies, effect sizes were calculated so that a positive number indicated that individual therapy was more effective and a negative number indicated that group therapy was superior. Although individual therapy tended to produce better results than group therapy in these five studies, the effect sizes representing the difference between the two methods (M = 0.31, SD = 0.35) did not differ reliably from zero, r(4) = 1.95, p = .1.

In studies comparing group therapy with a wait-list control group, the number of clients per therapy group ranged from 3 to 12 (M = 7). Some might suspect that smaller therapy groups would be more effective than larger groups. However, we detected no consistent relation when we correlated group size with the outcome of therapy, r(12) = −.13, p = .7. Thus, small treatment groups did not appear to be any more effective than therapy groups with more clients.

**Type of outcome.** A wide range of outcome measures was used in these studies. Some of the instruments were designed specifically to assess changes in depressive symptomatology. Others measured constructs that were less directly related to depression. To determine whether the effects of therapy varied across these different domains of outcome, effect sizes representing comparisons of treated clients with wait-list controls were calculated separately for measures of depression and measures of other constructs. These results are presented in the top half of Table 7. For both types of outcome, there was a reliable effect of therapy, with treated clients improving more than wait-list controls, but the difference between the effect sizes for the different measures was not reliable, r(47) = 1.37, p = .2. However, in the 20 studies that included both types of measures,
Table 7  
Efficacy of Psychotherapy for Different Types of Outcome Measures

<table>
<thead>
<tr>
<th>Measure characteristic</th>
<th>N of studies</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression</td>
<td>29</td>
<td>0.93*</td>
<td>0.76</td>
</tr>
<tr>
<td>Other</td>
<td>20</td>
<td>0.64*</td>
<td>0.63</td>
</tr>
<tr>
<td>Source</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-report</td>
<td>29</td>
<td>0.85*</td>
<td>0.73</td>
</tr>
<tr>
<td>Observer</td>
<td>7</td>
<td>0.81*</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Note. Means and standard deviations are based on weighted least squares analyses in which effect sizes were weighted by sample size. *p < .05.

There was a tendency for effect sizes to be larger for measures designed to assess depression (M = 0.85, SD = 0.79) than for other instruments (M = 0.64, SD = 0.63), paired t(19) = 1.81, p = .09.

Estimates of the effects of therapy might also vary with the perspective of the evaluator. In the studies under review, two sources of outcome were commonly used: client self-report and reports from independent observers. As indicated in the bottom half of Table 7, the effect of therapy was substantial from both perspectives and quite similar for clients and observers, t(34) = 0.13, p = .9. In seven studies that included both sources of outcome, effect sizes from observer measures (M = 0.81, SD = 0.77) were somewhat larger than those for self-report instruments (M = 0.70, SD = 0.96), but the difference was once again not reliable, paired t(6) = 0.56, p = .6.

Sample characteristics. The client samples used in these studies varied on a number of dimensions. For example, some studies used female clients exclusively, whereas others included both sexes. The mean age of the client sample also varied across studies. It may be that the effects of psychotherapy on depression differ depending on these client characteristics. However, analysis indicated that these demographic variables had no appreciable relation with treatment outcome. In comparisons between treated groups and wait-list controls, effect sizes did not vary reliably as a function of the percentage of female clients, r(24) = −.19, p = .3, or the age of the sample, r(19) = −.23, p = .3.

Initial severity of depression is another factor that might moderate the efficacy of treatment. One indicator of initial severity is the pretreatment score of clients on the BDI. When we examined comparisons of treated groups with wait-list controls, however, we did not find a systematic relation between effect size and initial scores on the BDI, r(24) = 0.25, p = .3. Thus, there was no convincing evidence that clients scoring higher on this measure responded differently to treatment than did those who initially reported less depression.

Another indicator of the severity of depression is the diagnostic status of the sample. To determine the impact of this variable, we divided studies into two groups: those that used formal diagnostic criteria in the screening process and those that used less rigorous client selection procedures. Effect sizes for comparisons of treated clients with wait-list controls are reported for both types of studies in the top half of Table 8. As this table reveals, the effects of therapy were virtually identical for formally diagnosed samples and for those that were not formally diagnosed, t(27) = 0.08, p > .9.

A final sample characteristic that might affect study outcome is the referral source of the clients. In most of the studies, clients were obtained through media announcements. Only rarely were traditional outpatient referral procedures used. To assess the influence of referral source, we compared the results of studies using outpatients with those of studies using student or community volunteers. Effect sizes for comparisons of treated clients with wait-list controls are presented for the three groups of studies in the bottom half of Table 8. As the values suggest, there was no reliable relation between effect size and referral source, F(2, 23) = 0.34, p = .7. Thus, treatment effects in studies using solicited clients did not differ from those observed in studies using traditionally referred outpatients.

Other characteristics of the studies. In addition to using different client samples, the reviewed studies also varied on a number of other characteristics. One such factor that can have a major impact on the validity of a study is the degree to which attrition rates differ for treatment and control groups. When the proportion of dropouts is unequal, the therapy manipulation becomes confounded with these differences in attrition. Thus, a treatment may appear ineffective simply because it has lost more of its successful cases. If this were the case, then one would expect studies with greater attrition in treatment groups than in control groups to show smaller treatment effects. To investigate this possibility, we created a differential attrition variable for each study, which was defined as the percentage of attrition from treatment groups minus the percentage of attrition from wait-list control conditions. When we correlated the treatment effect reported in a study with this differential attrition variable, no reliable relation emerged, r(20) = −.13, p = .6.

The level of training of therapists is another factor that might affect the outcome of a study. If student therapists are less skilled than professionals, studies using students might provide an inadequate test of treatment benefits. However, our results argue against this possibility. In fact, using comparisons of treated clients with wait-list controls, we found that the 19 studies in which graduate students served as therapists yielded larger effect sizes (M = 1.05, SD = 0.73) than those obtained from the

Table 8  
Efficacy of Psychotherapy for Different Client Samples

<table>
<thead>
<tr>
<th>Sample characteristic</th>
<th>N of studies</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formal diagnosis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>8</td>
<td>0.86*</td>
<td>0.85</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
<td>0.84*</td>
<td>0.64</td>
</tr>
<tr>
<td>Referral source</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solicited from community</td>
<td>13</td>
<td>0.75*</td>
<td>0.64</td>
</tr>
<tr>
<td>Solicited students</td>
<td>11</td>
<td>0.95*</td>
<td>0.82</td>
</tr>
<tr>
<td>Outpatients</td>
<td>2</td>
<td>1.13</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Note. Means and standard deviations are based on weighted least squares analyses in which effect sizes were weighted by sample size. *p < .05.
2 studies in which all therapists were fully trained professionals \((M = 0.65, SD = 0.05)\), although this numerical difference was not statistically significant, \(t(19) = 0.70, p = .5\).

**Analysis 2: Comparisons With Pharmacotherapy**

As our analyses indicate, psychotherapy appears effective in helping depressed clients. A further question, though, is the effectiveness of psychotherapy relative to the leading alternative treatment, pharmacotherapy. Several previous quantitative reviews have suggested that psychotherapy compares favorably with drug therapy for the treatment of depression (e.g., Quality Assurance Project, 1983; Smith et al., 1980; Steinbrueck et al., 1983). However, in each of these reviews the efficacy of psychotherapy and pharmacotherapy was estimated from different sets of studies. Thus, differences in variables such as client characteristics may account for the apparent variation in treatment effects. In addition, the benefits of combining psychotherapy with pharmacotherapy have received scant attention. Although several reviewers using narrative or box-score techniques have suggested that the combination approach may be more effective than either treatment alone (e.g., Conte et al., 1986; Weissman, 1979), the evidence has not been subjected to rigorous inferential analysis.

Fortunately, a number of studies have recently been published in which both psychotherapy and pharmacotherapy or the combination of the two have been evaluated for the treatment of depression. The following analyses provide a quantitative summary of this research.

**Method**

To locate appropriate studies, we conducted computerized searches of both medical and psychological data bases and examined the references of published reviews and outcome studies in this area. In addition, recent volumes of relevant journals were searched on an issue-by-issue basis. We selected studies containing at least one of the following comparisons: (a) psychotherapy versus pharmacotherapy, (b) a psychotherapy–pharmacotherapy combination versus psychotherapy alone, and (c) a combination treatment versus pharmacotherapy alone. Comparisons with “treatment as usual” were excluded when the treatment-as-usual condition could not be clearly classified as psychotherapy, pharmacotherapy, or the combination of the two. In addition, when a psychotherapy–pharmacotherapy combination was evaluated, we required that it be compared with one of its components. For example, a comparison between cognitive therapy and the combination of psychodynamic therapy and imipramine would have been excluded, because the difference between the two conditions reflects not only the addition of pharmacotherapy but also a difference in the type of psychotherapy. Other inclusion criteria paralleled those outlined earlier for the larger psychotherapy review. Thus, we included only studies using adult outpatients suffering from unipolar depression, and once again the focus was on treatments with individual or group formats.

Using these criteria we were able to locate 15 studies that provided the relevant comparisons between psychotherapy, pharmacotherapy, or the combination of the two. (Appendix B presents a list of the references for the studies.) Many of these studies were reported in more than one publication. In fact, on average more than two reports were published from each research project. Thus, the actual number of separate studies in this area is far smaller than might be expected on the basis of the number of publications.

The client samples used in these studies were similar to those of the psychotherapy studies reviewed earlier in terms of age \((M = 41.7\) years\), gender \((M = 68.5\% \) female\), and initial BDI score \((M = 25.5)\). In all except one of the studies, clients were required to meet formal diagnostic criteria for depression. Only three of these studies (21%) reported that clients were solicited through media announcements. Five studies (36%) used traditionally referred outpatients, and the remaining seven studies (47%) either included both solicited and traditionally referred clients or did not report the source of their subjects.

A variety of psychotherapeutic approaches were evaluated in this set of studies. Cognitive–behavioral therapy was used in eight investigations (53%), and behavioral treatments were tested in three studies (20%). One study (7%) examined purely cognitive methods, whereas four studies (27%) assessed a general verbal therapy. In one study, the type of therapy was not reported.

In 12 of the studies (80%), tricyclic antidepressants constituted the pharmacotherapy under investigation, with amitriptyline most often administered. One study (7%) examined a tetracyclic, and another 2 (13%) evaluated benzodiazepines. A final study allowed prescribing physicians to use any of a variety of psychoactive drugs. Mean dosages of these medications were generally within accepted therapeutic ranges, although usually at the lower end. Treatment lasted on average 12 weeks, allowing in most cases an adequate trial of the medication.

For each study, effect sizes representing differences in treatment outcome were calculated by using the procedures outlined for the earlier psychotherapy analyses. Once again, multiple outcomes within a treatment comparison and multiple treatment comparisons within a study were averaged to ensure that estimates of error were based on independent observations. As before, a weighted least squares procedure was used to compensate for differences in the sample sizes of the studies.

**Results**

The top half of Table 9 presents the results of comparisons between psychotherapy, pharmacotherapy, and the combination of both. As can be seen, psychotherapy appeared more effective than pharmacotherapy in the treatment of depression. However, the outcomes of a combination approach did not differ systematically from the outcomes of either treatment alone. Thus, the benefits of psychotherapy and pharmacotherapy do not appear to be additive.

Could the superiority of psychotherapy over pharmacotherapy be an artifact of researcher allegiance? To investigate this possibility, we had two independent raters judge researcher allegiance by using the procedures outlined in the preceding psychotherapy review. We then used these allegiance ratings to predict the effect size that would have occurred if the research-
Table 9  
**Relative Efficacy of Psychotherapy, Pharmacotherapy, and the Combination of Both**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>N of studies</th>
<th>Effect size*</th>
<th>Estimate if allegiance</th>
</tr>
</thead>
<tbody>
<tr>
<td>All pharmacological treatments</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychotherapy vs. drug therapy</td>
<td>8</td>
<td>0.13*</td>
<td>0.12 0.07 (0.04)</td>
</tr>
<tr>
<td>Combination vs. psychotherapy</td>
<td>12</td>
<td>0.01</td>
<td>-0.01 0.08 (0.08)</td>
</tr>
<tr>
<td>Combination vs. drug therapy</td>
<td>5</td>
<td>0.17</td>
<td>0.24 -0.05 (0.21)</td>
</tr>
<tr>
<td>Tricyclic antidepressants only</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychotherapy vs. tricyclics</td>
<td>7</td>
<td>0.12*</td>
<td>0.13 0.07 (0.04)</td>
</tr>
<tr>
<td>Combination vs. psychotherapy</td>
<td>9</td>
<td>0.02</td>
<td>0.25 -0.05 (0.08)</td>
</tr>
<tr>
<td>Combination vs. tricyclics</td>
<td>4</td>
<td>0.15</td>
<td>0.27 -0.05 (0.26)</td>
</tr>
</tbody>
</table>

*Note. Means, standard deviations, and standard errors are based on weighted least squares analyses in which effect sizes were weighted by sample size.

Method

We were able to identify a total of 39 studies reporting normative data on the BDI. (See Appendix C for a list of the references for the studies.) Of these studies, 20 had been previously identified by Nietzel et al. (1987) in their search of the 1978 through September 1985 issues of six psychology journals. We located another 19 normative studies by extending Nietzel et al.'s journal search from September 1985 through November 1987 and by reviewing bibliographies of articles on the assessment of depression. These normative studies included samples that were similar in terms of their age and gender to the clients treated in the psychotherapy studies. The average normative sample consisted of 64% female subjects, and the average age of those in the sample was 30 years.

In 28 of the normative studies, the BDI was administered to a group of subjects unscreened for mental health difficulties. Examples include investigations that randomly sampled community residents and those using data from large groups of university students. In 12 studies, norms were reported for individuals who had been screened on some measure of mental health. An example is a study that obtained data from subjects who did not meet Research Diagnostic Criteria for depression. Not surprisingly, BDI scores for nondistressed samples were lower than those derived from samples of the general population, t(38) = 2.14, p = .04. For this reason, we calculated norms separately for nondistressed and general population samples.

Results

Table 10 presents BDI scores for both types of norms along with pretreatment and posttreatment scores derived from 22 studies that compared clients who received psychotherapy with no-treatment controls. As can be seen, both treated clients and untreated controls were moderately depressed at pretest. After treatment, clients receiving psychotherapy were functioning at the lower end of the mild range of depression, whereas controls scored at the upper end of the mildly depressed range. Despite this improvement relative to untreated controls, clients who had received psychotherapy remained more depressed at the end of therapy than subjects in the general population studies, t(23.7) = 4.95, p < .001. A similar picture emerged when the more stringent norms for the nondistressed samples were used. Once again, treated clients were more depressed at the end of therapy than subjects in the normative studies, t(31.4) = 6.23, p < .001. Thus, although psychotherapy brings about a reduction in depression, clients who have received therapy can still...
Table 10
Beck Depression Inventory (BDI) Scores for Normative Samples, Clients Treated With Psychotherapy, and Untreated Controls

<table>
<thead>
<tr>
<th>Group</th>
<th>N of studies</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normative samples</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>General population</td>
<td>28</td>
<td>7.0</td>
<td>1.3</td>
</tr>
<tr>
<td>Nondistressed</td>
<td>12</td>
<td>4.9</td>
<td>2.0</td>
</tr>
<tr>
<td>Treated clients</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretreatment</td>
<td>22</td>
<td>21.8</td>
<td>4.7</td>
</tr>
<tr>
<td>Posttreatment</td>
<td>22</td>
<td>11.8</td>
<td>4.4</td>
</tr>
<tr>
<td>Untreated controls</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretreatment</td>
<td>22</td>
<td>20.7</td>
<td>5.0</td>
</tr>
<tr>
<td>Posttreatment</td>
<td>22</td>
<td>18.1</td>
<td>5.2</td>
</tr>
</tbody>
</table>

Note. Means and standard deviations are based on weighted least squares analyses in which BDI scores were weighted by sample size.

BDI* scores from 10 to 20 indicate mild depression; from 20 to 30, moderate depression; and over 30, severe depression (Kendall, Hollon, Beck, Hammen, & Ingram, 1987).

be distinguished from healthy individuals who have not sought treatment.

Jacobson et al. (1984) have suggested that one way to judge the impact of treatment is to express client functioning relative to the mean and standard deviation of a population of healthy normal individuals. To generate such estimates, we pooled the standard deviations of BDI scores reported in studies of the general population and those reported in studies of nondepressed samples, and we used the pooled values as estimates of the dispersion of individual scores in each normative population. We then expressed the BDI scores of clients treated with psychotherapy in terms of these normative standard deviations. The results, presented in Figure 1, indicate that before treatment clients reported substantially more depression on the BDI than did normative subjects. After receiving psychotherapy, however, clients had moved to within one standard deviation of the mean of the general population and within one and a half standard deviations of the mean of nondistressed samples.

Discussion
As our analyses demonstrate, clinical research has firmly established the efficacy of psychological interventions for depression. In studies comparing psychotherapy with no treatment, there was a positive effect of therapy that was both statistically reliable and substantial in magnitude. Despite the relatively brief duration of treatment provided in most of the reviewed studies, the average client was functioning at treatment termination approximately three fourths of a standard deviation better than untreated controls.

It remains unclear, though, which aspects of psychotherapeutic treatment were responsible for producing this improvement. When the effects of psychotherapy were compared with those of placebo treatments, no reliable differences emerged. Such results are somewhat surprising, because previous quantitative reviews of the psychotherapy outcome literature have generally shown that the effects of therapy are larger than those of placebo treatments (e.g., Bowers & Clum, 1988; Casey & Berman, 1985; Dush et al., 1983; Miller & Berman, 1983; Shapiro & Shapiro, 1983). However, these previous reviews were not restricted to depressed samples. It may be that depression is particularly responsive to common curative factors occurring in both psychotherapy and placebo treatments. If so, then the specific procedures that define the type of treatment may be less important in alleviating depression than previously recognized.

Despite their improvement, clients treated with psychotherapy remain distinguishable from healthy controls. When we compared the BDI scores of treated clients with norms derived from the general population and from nondistressed samples, treated clients were found to be more depressed at the end of therapy than both normative groups. However, the magnitude of improvement over the course of therapy appeared impressive. On average, clients were functioning within one standard deviation of the general population after treatment, compared with a pretreatment difference of more than two standard deviations. Such a change clearly represents substantial improvement.

The research evidence demonstrates, furthermore, that the benefits of psychotherapy for depression are not short-lived. In those studies that included a follow-up assessment, improvement at posttreatment was quite similar to that observed at a later follow-up. Not only does this finding emphasize the enduring nature of the changes that occur during treatment, but it

![Figure 1](https://example.com/Figure1.png)
argues against the widely held assumption that follow-up assessment is essential because of unusually high relapse rates for depression. As in the more general review by Nicholson and Berman (1983), our results suggest that follow-up findings often add little to information obtained at the end of therapy. Thus, rather than including a costly follow-up assessment, psychotherapy researchers might justifiably choose to invest resources in other aspects of their design. For example, researchers could focus on obtaining larger samples of clients and therapists, thereby increasing both the power and generalizability of their analyses.

Although all forms of psychotherapy were more effective than no treatment, our initial analyses indicated that there might be some variation in the efficacy of different types of therapy. Such a difference has also been reported in a recent review by Dobson (1989), who concluded that cognitive therapy for depression was superior to other approaches. However, after more detailed analyses, we discovered that differences in the efficacy of different treatments may be an artifact of the theoretical allegiances of the researchers conducting these studies. When a particular type of therapy was preferred by an investigator, it tended to produce more favorable results than the treatment with which it was being compared. This pattern is similar to findings reported by Berman et al. (1985) and Smith et al. (1980, chap. 5), who also found that the outcome of a psychotherapy study varied according to the allegiance of the researcher. Furthermore, the effect is consistent with Rosenthal’s (1969, 1976) well-known research demonstrating the influence of experimenter expectations. To take this influence into account, we used regression analysis to predict the outcomes that would have occurred under conditions of no allegiance. This predictive analysis indicated that if all of the researchers had been neutral, there would have been no reliable differences in the effectiveness of the various types of therapy.

The allegiance of the investigator also appeared to play a role in comparisons between psychological and pharmacological treatments for depression. Although psychotherapy initially appeared more effective than drugs, this difference was not reliable once we adjusted for the influence of investigator allegiance. Our analyses also indicated that combinations of psychotherapy with pharmacotherapy were not systematically more effective than either of the treatments alone. Furthermore, the efficacy of psychotherapy and pharmacotherapy remained comparable even when we restricted the analysis to studies using tricyclic antidepressants, the best established of the pharmacological interventions for depression.

Advocates of drug therapy might argue that the effectiveness of pharmacotherapy was underestimated in these studies for a variety of reasons. For example, the clinical response to pharmacotherapy might have been better had higher levels of the drugs been prescribed. In addition, some have suggested that the efficacy of pharmacological treatment depends partly on client characteristics such as chronicity and the presence of endogenous symptoms (e.g., Becker & Schuckit, 1978; Weissman, 1981). Thus, given different dosages or client samples, pharmacotherapy could have appeared more effective. Alternatively, advocates of particular psychological or pharmacological approaches may argue that our classification system was overly broad, thereby obscuring important differences in the efficacy of specific forms of treatment. Although such arguments have merit, the onus would appear to be on the proponents of such hypotheses to offer convincing empirical evidence in support of them.

The relative efficacy of group and individual therapy has only recently become an issue in depression research, perhaps because group therapy for depression has traditionally been considered a contradiction in terms. It has been feared that the special needs of depressed clients could not be met in a group setting and that depressed clients might act as a burden on a group because of their withdrawal, pessimism, and self-absorption. However, our analyses indicated that both group and individual treatment formats produced more improvement than no treatment, and the effects of the two approaches were comparable. Moreover, there was no evidence that treatment groups with many members were less effective than smaller therapy groups. Although this lack of difference in the efficacy of individual and group treatments is consistent with the findings of several previous reviews (Miller & Berman, 1983; Shapiro & Shapiro, 1982; Smith et al., 1980, chap. 5), it conflicts with Nietzel et al.’s (1987) and Dush et al.’s (1983) evidence for the superiority of individual therapy. The difference between our results and those of Nietzel et al. are particularly striking, because their review also focused specifically on the treatment of depression. However, it should be noted that our reviews differ not only in terms of the studies included but also in the outcome measures analyzed.

Although variations in sample characteristics such as age, sex, and initial symptom severity have often been offered as explanations for inconsistencies in the results of different studies, we found no evidence that any of these variables were systematically related to outcome. Furthermore, studies that included only clients meeting formal diagnostic criteria for depression generated results that were virtually identical to those observed in studies using less rigorous client selection procedures. The reliance on self-report measures for client selection has often been criticized on the grounds that these measures were not developed as diagnostic tools. However, our data suggest that, at least for the problem of depression, treatment findings based on self-identified clients will mirror the results obtained from formally diagnosed samples.

Self-report instruments have been viewed as suspect not only when used for client selection but also when used as an outcome measure. The underlying assumption is that clients may overestimate the benefits of treatment, whereas measures obtained from observers will exhibit less bias. However, in a recent meta-analysis, Lambert, Hatch, Kingston, and Edwards (1986) found greater indications of change on an interviewer measure of depression than on two self-report scales. In our analysis of a broad range of self-report and observer measures, we found that treatment effects derived from self-report instruments were comparable in size to those obtained from observer measures. As in the Lambert et al. (1986) review, there was no evidence that measures based on the self-report of the client yielded overly positive estimates of treatment efficacy.

Our analysis did suggest that the effects of therapy differed according to the content of the outcome measures. Instruments specifically designed to assess depression tended to produce larger effects than measures assessing other constructs. Such
findings might suggest that the psychotherapies used in these
studies are more effective for depression than for other types
of complaints. However, it must be remembered that we only
included studies with clients whose primary problem was de-
pression. Therefore, an equally plausible explanation for these
results is that change is more likely to occur for the problems
or symptoms that initially prompt a client to enter treatment.

In our review, we included studies that varied considerably in
terms of the quality of their research design. Critics of quantita-
tive review methods (e.g., Wilson, 1985; Wilson & Rachman,
1983) have objected to this practice, arguing that the quality of
the research should be taken into account in the selection of
studies. Like other quantitative reviewers (e.g., Glass et al.,
1981, chap. 2), we consider the impact of study quality to be an
empirical question. In fact, our analyses generally failed to de-
tect differences as a function of research design characteristics.
For instance, findings from studies with differential rates of at-
trition did not differ reliably from the findings of studies in
which dropout rates were equivalent across treatment groups.
Moreover, consistent with the findings of Berman and Norton
(1985), investigations that used fully trained professionals as
therapists did not report greater treatment benefits than studies
that relied on graduate students to administer therapy. Further-
more, studies that used treatment manuals or that monitored
therapist behavior during treatment yielded findings similar to
those of studies that did not include these procedures for ensur-
ing correct treatment delivery.

Of the various study characteristics examined, only the num-
ber of clients per study related to treatment outcome. As our
preliminary analysis of the psychotherapy data indicated, stud-
ies with fewer clients yielded larger effects than studies with
many clients. The most likely explanation for this relation is
that it reflects a publication bias. That is, journals are more
likely to publish studies that achieve statistically significant
differences between groups, and small studies need larger effects
to achieve statistical significance. One implication of this find-
ing is that literature reviews of published research (whether they
are narrative summaries or quantitative reviews such as our
own) may yield overly generous estimates of treatment effects.

Because of the way in which we conducted our analyses, how-
ever, such a publication bias was probably minimized in this
review. For example, our statistical analyses gave less weight to
studies with smaller samples, and it is these small-sample stud-
ies that are most likely to inflate estimates of treatment effects.
In addition, our procedures for estimating treatment effects
were often conservative. Thus, when results were reported sim-
ply as nonsignificant, we estimated that effect size to be zero.
Whether or not these conservative estimation procedures fully
offset the influence of a publication bias, our review can at least
be viewed as a conservative summary of the published evidence
on the efficacy of psychotherapy for depression.

The findings in this review are consistent with more general
quantitative reviews of the psychotherapy outcome literature
(e.g., Smith et al., 1980). As with these other reviews, we found
few differences in the efficacy of various therapeutic methods.
However, unlike these general analyses, ours provided an eval-
uation of comparative treatment efficacy within a single diagnos-
tic category. Thus, the similarity in the effects of various thera-
pies cannot be attributed to their application to different noso-
logical groups. Even when the client sample was restricted to
the diagnosis of depression, different types of psychotherapy
yielded equivalent benefits. Furthermore, the similarity be-
tween our findings and those of general reviews including other
diagnostic groups raises the provocative possibility that diagnos-
isis has much less impact on treatment response than previously
assumed. Psychotherapy may produce similar benefits not only
across different types of therapy but also across different types
of clients. Thus, extensive efforts to establish the diagnostic pu-
rity of samples for psychotherapy research may be unnecessary.

The evidence from the review, coupled with that of previous
quantitative reviews, indicates that additional comparisons be-
tween competing therapies for depression are not likely to prove
informative. Comparative studies are useful primarily in estab-
lishing new approaches as viable therapeutic options rather
than in elucidating the mechanisms through which they effect
treatment. As Kazdin (1981) has noted in a commentary on de-
pression outcome research: “Relatively little is known about
major treatment contenders, including cognitive therapy and
behavior therapy, and the comparative studies do not shed light
on whether these individual treatments operate in many of the
ways proposed on conceptual grounds” (p. 319). Indeed, the
key question now is not whether psychotherapy works for the
treatment of depression but rather how these therapies produce
their benefits.

In a recent analysis of the cognitive theory of depression, Hol-
on, DeRubeis, and Evans (1987) pointed out that the similarity
in the effects of differing therapies does not necessarily disprove
the causal significance of cognitive changes as a mediator of re-
covery from depression. The possibility remains that all thera-
pies are effective because all of them activate the cognitive
changes that are the specific target of cognitive therapy. Al-
though this argument could be true, it must be recognized that
a comparable position could be advanced with equal plausibil-
ity by proponents of behavior therapy. That is, therapies may
promote recovery by producing behavioral changes, whether by
accident or by design. In this sense, cognitive or behavioral
changes may constitute nonspecific effects that occur naturally
over the course of therapy regardless of whether they are the
identified goal of treatment. Alternatively, factors such as cli-
ents’ expectations of improvement, their acceptance of the ther-
apeutic rationale, or the quality of the therapeutic relationship
may be the central mechanisms through which therapeutic
change occurs (e.g., see Frank, 1982; Zeiss, Lewinsohn, & Mu-
noz, 1979).

Surprisingly little research has been directed toward evaluat-
ing these hypothesized mediators of change. Curative factors
common to all therapies have generally been mentioned only as
post hoc explanations in comparative studies finding nonsig-
nificant differences in treatment outcomes. If researchers are to
progress in their understanding of how psychotherapy benefits
clients, these common factors may need to become a more cen-
tral focus of future research efforts.

References

American Psychiatric Association. (1980). Diagnostic and Statistical

Atkinson, A. K., & Rickel, A. U. (1984). Postpartum depression in pri-


Kilpatrick-Tabak, B., & Roth, S. (1978). An attempt to reverse perfor-


Appendix A

Sources for the Psychotherapy Studies Included in the Review

Information about a study was sometimes obtained from more than one published report. In those cases, the other reports providing information about the study are listed in parentheses. Also indicated in parentheses are those cases in which a report provided information about more than one study.


DiMascio, A., Weissman, M. M., Prusoff, B. A., Neu, C., Zwilling, M., & Klerman, G. L. (1979). Differential symptom reduction by drugs and psychotherapy in acute depression. *Archives of General Psychiatry, 36*, 1450–1456. (Information about this study was also obtained from Herceg-Baron et al., 1979; Prusoff et al., 1980; Weissman et al., 1981; Weissman, Prusoff, & DiMascio, 1979; Weissman, Prusoff, DiMascio et al., 1979; Zuckerman et al., 1980.)


Hammen, C. L., & Glass, D. R. (1975). Depression, activity, and evaluation of reinforcement. *Journal of Abnormal Psychology, 84*, 718–721. (This report provided information about two different experiments. Each of these experiments was treated as a separate study in the analyses.)


Weissman, M. M., Prusoff, B. A., & Klerman, G. L. (1982). Depressed outpatients: Results one year after treatment with drugs and/or interpersonal psychotherapy. *Archives of General Psychiatry, 39*, 51-55. (Information about this study was also obtained from DiMascio et al., 1979; Herceg-Baron et al., 1979; Prusoff et al., 1980; Weissman, Prusoff, & DiMascio, 1979; Weissman, Prusoff, DiMascio et al., 1979; Zuckerman et al., 1980.)


Appendix B

Sources for the Pharmacotherapy Studies Included in the Review

Information about a study was sometimes obtained from more than one published report. In those cases, the other reports providing information about the study are listed in parentheses.


Bellack, A. S., Hersen, M., & Himmelhoch, J. (1981). Social skills training compared with pharmacotherapy and psychotherapy in the treatment of unipolar depression. *American Journal of Psychiatry, 138,* 1562−1567. (Information about this study was also obtained from Bellack et al., 1983a; 1983b; Hersen et al., 1984.)

Bellack, A. S., Hersen, M., & Himmelhoch, J. M. (1983a). A comparison of social-skills training, pharmacotherapy, and psychotherapy for depression. *Behaviour Research and Therapy, 21,* 101−107. (Information about this study was also obtained from Bellack et al., 1981, 1983b; Hersen et al., 1984.)


Blackburn, I. M., & Bishop, S. (1981). Is there an alternative to drugs in the treatment of depressed ambulatory patients? *Behavioural Psychotherapy, 9,* 96−104. (Information about this study was also obtained from Blackburn & Bishop, 1983; Blackburn et al., 1981, 1986.)

Blackburn, I. M., & Bishop, S. (1983). Changes in cognition with pharmacotherapy and cognitive therapy. *British Journal of Psychiatry, 143,* 609−617. (Information about this study was also obtained from Blackburn & Bishop, 1981; Blackburn et al., 1981, 1986.)


Blackburn, I. M., Eunson, K. M., & Bishop, S. (1986). A two-year naturalistic follow-up of depressed patients treated with cognitive therapy, pharmacotherapy and a combination of both. *Journal of Affective Disorders, 10,* 67−75. (Information about this study was also obtained from Blackburn & Bishop, 1981; Blackburn et al., 1981.)


DiMascio, A., Weissman, M. M., Prusoff, B. A., Neu, C., Zwilling, M., & Klerman, G. L. (1979). Differential symptom reduction by drugs and psychotherapy in acute depression. *Archives of General Psychiatry, 36,* 1450−1456. (Information about this study was also obtained from DiMascio, Klerman et al., 1979; Hercog-Baron et al., 1979; Prusoff et al., 1980; Rounsaville et al., 1981; Weissman et al., 1981; Weissman, Prusoff, & DiMascio, 1979; Weissman, Prusoff, DiMascio et al., 1979; Zuckerman et al., 1980.)


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. (Information about this study was also obtained from Simons, Garfield, & Murphy, 1984; Simons, Levine et al., 1984; Simons et al., 1985, 1986.)

Prusoff, B. A., Weissman, M. M., Klerman, G. L., & Rounsaville, B. J. (1980). Research diagnostic criteria subtypes of depression: Their role as predictors of differential response to psychotherapy and drug treatment. *Archives of General Psychiatry, 37,* 796−801. (Information about this study was also obtained from DiMascio, Klerman et al., 1979; DiMascio, Weissman et al., 1979; Hercog-Baron et al., 1979; Rounsaville et al., 1981; Weissman et al., 1981; Weissman, Prusoff, & DiMascio, 1979; Weissman, Prusoff, DiMascio et al., 1979; Zuckerman et al., 1980.)


Rounsaville, B. J., Klerman, G. L., & Weissman, M. M. (1981). Do psychotherapy and pharmacotherapy for depression conflict? *Archives of General Psychiatry, 38,* 24−29. (Information about this study was also obtained from DiMascio, Klerman et al., 1979; DiMascio, Weissman et al., 1979; Hercog-Baron et al., 1979; Prusoff et al., 1980; Weissman et al., 1981; Weissman, Prusoff, & DiMascio, 1979; Weissman, Prusoff, DiMascio et al., 1979; Zuckerman et al., 1980.)


Appendix C

Sources for the Normative Studies Included in the Review


Received August 12, 1988
Revision received June 27, 1989
Accepted August 10, 1989