coming, but in 1975 Luborsky, Singer, and Luborsky examined about 100 comparative treatment studies and found that Rosenzweig’s hypothesis was essentially right: There was a trend of only relatively small differences from comparisons of outcomes of different treatments. Around that time researchers began to call such small differences by the title from Rosenzweig’s quote from *Alice in Wonderland*: “everybody has won, so all shall have prizes” which was the “Dodo bird’s verdict” after judging the race. The term “Dodo bird verdict” has since become commonly used, and researchers have continued to write articles for or against the existence of, or the meaning of, that trend.

In this study we aimed to survey and then to evaluate whether Rosenzweig (1936) hypothesis is still fitting and still flourishing. We examined the exact amount of support for this trend, for the task is still very necessary; even expert psychotherapy researchers have different opinions, and even high effect, about the expected results. For a brief sample of these many opinions, see Beutler (1991); Crits-Christoph (1997); Cuijpers (1998); Henry (1998); Howard, Krause, Saunders, and Kopta (1997); Luborsky (1995); King (1997); King and Ollendick (1998); Luborsky, Diguer, Luborsky, and Schmidt (1999); Nietzel, Russell, Hemmings, and Gretter (1987); Reid (1997); Tschuschke et al. (1998); Wampold, Mondin, Moody, and Ahn (1997); and Wampold, Mondin, Moody, Stich, Benson, and Ahn (1997).

THE USE OF A COLLECTION OF META-ANALYSES TO CHECK THE DODO BIRD VERDICT

Before we present results of our collection of meta-analyses of studies on this topic we must review our rea-
sons for using them in the way we did. First, our collection only relies on meta-analyses because, according to Rosenthal and Rubin (1985) and Rosenthal (1998), meta-analyses ordinarily take into account each study’s sample size and the magnitude of effect size of the treatments compared in each study. An effect size is a type of measure of the degree of association of two variables. The measures in each of these meta-analyses is a Cohen’s d (or a variant of it), a difference between the two treatment means relative to their within-group variations.

Second, we have limited ourselves to meta-analyses of the relative efficacy of pairs of different active psychotherapies in comparison with each other. We chose to emphasize such comparisons of active treatments because (a) these give an assessment of the relative efficacy of different active treatments that is of greater interest to clinicians than comparisons of treatment versus controls, and (b) the background variables for the patients in each treatment are likely to be more comparable when there is a direct comparison of two different active types of psychotherapy. We will not deal here with the level of efficacy of these treatments, for there is much evidence already for their mostly good level of efficacy (e.g., Lambert & Bergin, 1994; Lipsy & Wilson, 1993; Shadish et al., 1997).

Third, we further limited our domain to meta-analyses of studies of common psychiatric diagnoses applied to adults (age 18 or older): depression, anxiety disorders (including obsessive-compulsive disorder and phobia), and mixed neurotics. Our findings do not apply to patients who are psychotic, nor do they apply to children.

Fourth, we also limited the scope of our review to meta-analyses of some of the common types of therapies: behavior therapy, cognitive therapy, cognitive-behavior therapy, dynamic therapy, rational-emotive therapy, and drug therapy. Drug therapy is included because it is one of the most common comparisons with psychotherapies, and psychotherapists tend to be especially interested in the results of this comparison.

The meta-analyses that fit our criteria are briefly discussed below. Our search for these meta-analyses was aided by a computer-based literature search (using the PsychInfo and Medline databases) and by two large lists of meta-analyses in Lipsy and Wilson (1993) and Chambless et al. (1996). For finding the meta-analyses in the computer sources, our search labels included “comparative treatment studies,” “nonsignificant difference effect,” “Dodo bird verdict,” and “empirically validated treatments.”

Comparisons of Effect Sizes of Active Treatments with Each Other

Rosenzweig (1936) reasoned that psychotherapy outcome studies would show that different psychotherapies seem to have major ingredients in common that would lead them to have only small and nonsignificant outcome differences; one such major ingredient is that they all involve a helping relationship with the therapist. Rosenzweig’s conclusion was confirmed by Luborsky, Singer, and Luborsky in 1975, as noted earlier.

In 1980 Smith, Glass, and Miller supported Luborsky et al.’s (1975) conclusion but by a more systematic and much larger review of 475 comparative treatment studies of psychotherapy. They found an average effect size (ES) of treatment versus control studies of psychotherapy of .85. The ES measure used in this study and in the present study are all variants of Cohen’s d, as described by Rosenthal (1991). A Cohen’s d of .85 can easily be interpreted as a difference between the two group means of 85% of the standard deviation. Note that Smith et al. did not present effect sizes as we did from comparisons of active treatments with other active treatments, but rather effect sizes for each type of therapy compared with controls.

Fortunately, in the early phase of the present review, we also surmised a likely weakness of the method of relying on a treatment versus a control as compared with the method of an active treatment versus another active treatment. This weakness is, in part, that the active treatment versus the control treatment tended to deal less well with the match of the background factors in the patients in the treatments compared. As an example, in Smith, Glass, and Miller (1980), the patients in the sample of studies in cognitive therapy may well have been less psychiatrically severe in their disorders than those who were in dynamic treatment. An active treatment versus another active treatment comparison might have equalized the groups of patients by the probable effects of randomization into each treatment. For this reason also we decided to restrict our sample of meta-analyses to only those that relied on the comparison of two active treatments. We have located 17 of such meta-analyses in 6 reports (Table 1). Each of these 6 meta-analytic reports is briefly described below.

1. Berman, Miller, and Massman (1985), with a larger and more inclusive sample of studies than Miller and Berman (1993), also found small and non-significant differences between cognitive therapy and desensitization (N =
ment comparisons, one of which reported a significant difference between dynamic versus cognitive-behavioral with a significant correlation of $-0.47$ ($N = 14$ studies).

4. Crits-Christoph (1992) found nonsignificant differences in effect sizes of comparisons of active treatments for dynamic versus other psychotherapies ($N = 11$ studies).

5. Luborsky, Diguer, Luborsky, Singer, and Dickter (1993), in a sample of three studies, again showed nonsignificant effect sizes for dynamic versus other psychotherapies. (Note: To avoid duplication, the studies that were the same as those in Luborsky, Diguer, Seligman, et al. [1999] were omitted here.)

6. Luborsky, Diguer, Seligman, et al. (1999) found nonsignificantly different effect sizes in comparisons of

---

### Table 1. Meta-analyses of comparisons of effect sizes of active treatments (by Cohen’s $d$)

<table>
<thead>
<tr>
<th>Reports</th>
<th>Meta-analyses ($n = 17$)</th>
<th>No. of Studies</th>
<th>Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Berman, Miller &amp; Massman, 1985$^{b}$ Cognitive vs. desensitization</td>
<td>20</td>
<td>.06</td>
<td></td>
</tr>
<tr>
<td>2. Robinson, Berman et al., 1990 Cognitive vs. behavioral</td>
<td>12</td>
<td>.12</td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td>Cognitive vs. C-B</td>
<td>4</td>
<td>.03</td>
</tr>
<tr>
<td>4.</td>
<td>Cognitive vs. general verbal$^c$</td>
<td>7</td>
<td>.47*</td>
</tr>
<tr>
<td>5.</td>
<td>Behavioral vs. C-B</td>
<td>8</td>
<td>.24*</td>
</tr>
<tr>
<td>6.</td>
<td>Behavioral vs. general verbal</td>
<td>14</td>
<td>.27*</td>
</tr>
<tr>
<td>7.</td>
<td>C-B vs. general verbal</td>
<td>8</td>
<td>.37*</td>
</tr>
<tr>
<td>8.</td>
<td>Dynamic$^d$ vs. C-B</td>
<td>6</td>
<td>.47*</td>
</tr>
<tr>
<td>9.</td>
<td>Dynamic$^d$ vs. behavioral</td>
<td>5</td>
<td>.10</td>
</tr>
<tr>
<td>10.</td>
<td>Dynamic$^d$ vs. nonspecific</td>
<td>3</td>
<td>.29</td>
</tr>
<tr>
<td>11.</td>
<td>Dynamic$^d$ vs. psychiatric treatment</td>
<td>5</td>
<td>.32</td>
</tr>
<tr>
<td>12.</td>
<td>Dynamic$^d$ vs. psychiatric treatment</td>
<td>6</td>
<td>.05</td>
</tr>
<tr>
<td>13.</td>
<td>Svartberg &amp; Stiles, 1991$^e$ Dynamic$^e$ vs. other</td>
<td>3</td>
<td>.00</td>
</tr>
<tr>
<td>14.</td>
<td>Cognitive vs. behavioral</td>
<td>9</td>
<td>.21</td>
</tr>
<tr>
<td>15.</td>
<td>Dynamic$^f$ vs. behavioral</td>
<td>7</td>
<td>.03</td>
</tr>
<tr>
<td>16.</td>
<td>Dynamic$^f$ vs. cognitive</td>
<td>4</td>
<td>.02</td>
</tr>
<tr>
<td>17.</td>
<td>Pharmacotherapy vs. psychotherapy</td>
<td>9</td>
<td>.41</td>
</tr>
</tbody>
</table>

Mean effect size (absolute value) .20 ($n = 17$) .12 ($n = 11$) weighted: .21$^g$ weighted: .14

Median effect size (absolute value) .21 .14 weighted: .21$^g$ weighted: .15

Note: C-B, cognitive-behavioral therapy.

$^a$Corrected for researcher’s allegiance (by a mean of the three measures; Luborsky, Diguer, Seligman, et al., 1999); study 5 was corrected for quality of research design.

$^b$Includes Miller and Berman (1983).

$^c$General verbal therapy comprises treatments such as psychodynamic, client-centered, and other forms of interpersonal therapy. These treatments have in common a relatively greater emphasis on insight rather than on the acquisition of a set of specific skills.

$^d$The original Cohen’s $r$ for these studies was converted to Cohen’s $d$ (Cohen, 1977).

$^e$Short-term psychodynamic psychotherapy.

$^f$Brief dynamic psychotherapy.

$^g$A variety of dynamic treatments.

$^h$Effect sizes weighted by sample size of each corresponding study.

$^i$Weighted, as described by Rosenthal and DiMatteo (in press) and Rosenthal, Hiller, Bornstein, Berry, and Brunell-Neuleib (in press).

$p < .05$.

---

20 studies). (A positive Cohen’s $d$ means only that the first treatment is more effective than the other treatment.)

2. Robinson, Berman, and Neimeyer (1990) reported six meta-analyses with four of them significant using uncorrected effect sizes ($N = 53$ studies). These imply that behavioral treatment is less effective than cognitive-behavioral therapy ($-0.24$); that cognitive-behavioral therapy is more effective than general verbal ($0.37$); that cognitive therapy is more effective than general verbal ($0.47$); and that behavioral treatment is more effective than verbal ($0.27$). But when the effect sizes are corrected for the researchers’ allegiance (by a method to be described below), they become lower and nonsignificant.

3. Svartberg and Stiles (1991) continued the search for relatively efficacious therapies by meta-analyses of treatment comparisons, one of which reported a significant difference between dynamic versus cognitive-behavioral with a significant correlation of $-0.47$ ($N = 14$ studies).
cognitive versus behavioral, dynamic versus behavioral, dynamic versus cognitive, and pharmaco-therapy versus psychotherapy (N = 29 studies). The last of these comparisons was included because the pair is often considered by clinicians as a usable option either singly or in combination.

THE MAIN TRENDS IMPLIED BY THE META-ANALYSES

The mean effect sizes in the 17 meta-analyses showed low and nonsignificant differences. The mean absolute value of the uncorrected Cohen’s \( d \) effect size for the 17 meta-analyses listed in Table 1 was .20, which is not large and is nonsignificant, given that it is the mean of absolute values. To calculate each of the effect sizes, we first converted the three that used Pearson’s \( r \) into Cohen’s \( d \) so that they were all expressed in terms of Cohen’s \( d \).

The effect sizes were further reduced after corrections for the researcher’s allegiance and for other factors. The researcher’s allegiance effect is another major influence that can alter the typically modest and nonsignificant difference effect. This effect is the association of measures of the researcher’s allegiance to each of the treatments compared with measures of the outcomes of the treatments. There had been hints of this effect for many years, as first noted by Luborsky et al. (1975). Now there is an exhaustive review of the topic (Luborsky, Diguer, Seligman, et al., 1999) that shows a well-established researcher’s allegiance effect: The correlation between the mean of 3 measures of the researcher’s allegiance and the outcome of the treatments compared was a huge Pearson’s \( r \) of .85 for a sample of 29 comparative treatment studies. The three measures, described in Luborsky, Diguer, Seligman, et al., are ratings of the reprint, ratings by colleagues who know the researcher’s work well, and self-ratings of allegiance by the researcher themselves.

This high correlation of the mean of the three allegiance measures with the outcomes of the treatments compared implies that the usual comparison of psychotherapies has a limited validity because so far it is not easy to rule out the presence of the large researcher allegiance effect. To make matters worse, it is not clear how the allegiance effect comes about. A variety of methods have been suggested by Luborsky, Diguer, Seligman, et al. (1999) for reducing the intrusion of the researcher’s allegiance, but, even when these methods are implemented, the impact of such methods is likely to remain ambiguous in the precise amount of correction to be applied. Among the recommended precautionary steps, it might be valuable to (a) include researchers with a variety of allegiances in the research group carrying out the study and (b) choose as a comparison to the preferred treatment a treatment that is equally likely to be judged as credible (Berman and Luborsky, in preparation; Berman and Weaver, 1997).

A sample of the effects of corrections are noted as follows: When the uncorrected correlations in Robinson et al. (1990) were corrected for researchers’ allegiance by the mean of their three corrected allegiance scores (the most common type of correction used here) (Luborsky, Diguer, Seligman, et al., 1999), the correlations become lower and nonsignificant. The data from Smith et al. (1980) were corrected for reactivity (meaning influence of the therapist or researcher) and Luborsky et al. (1993) data were corrected for the quality of the research design (Luborsky, Diguer, Seligman, et al., 1999). The more exact changes can be seen by comparing the uncorrected with the corrected effect sizes in Table 1. For example, in Luborsky, Diguer, Luborsky, Singer, and Dickter (1993), the uncorrected comparison of two active treatments effect size was .00 (nonsignificant), and the effect size after correcting for research quality was similar in size (−.01; nonsignificant).

To summarize these results, we compared the mean of the effect sizes of corrected comparisons of active treatments from 11 meta-analyses in Table 1 (the 11 were all those for which we had data to compute corrections) with the mean of the corresponding uncorrected effect sizes. We first converted all these effect sizes into Cohen’s \( d \) (Cohen, 1977) and then took the mean of the absolute value of the effect sizes. The mean uncorrected effect size with Cohen’s \( d \) was .20, but the mean corrected Cohen’s \( d \) effect size was only .12; the reductions of the corrected effect sizes meant they were no longer significant. Also, the median uncorrected effect size was .21 as compared to a corrected median effect size of .14; the reductions also meant they were no longer significant.

Comparison of effect sizes of meta-analyses for cognitive and cognitive-behavioral versus dynamic and other treatments yielded small differences. A few of the common types of comparisons among the 17 meta-analyses warrant an even more focused review. But first, to make comparisons easier, two related subclasses can be combined—the cognitive and the cognitive-behavioral. It may be reasonable to combine these subclasses because
their comparison yielded only a very small effect size of −.03 in one meta-analysis with 4 studies (Robinson et al., 1990). An example of a comparison of corrected effect sizes that is easily shown in Table 1 compares cognitive or cognitive-behavioral treatment with other treatments. The mean effect size of these six comparisons of Cohen’s d is .14 (cognitive vs. behavioral, .12, cognitive vs. general verbal, −.15, behavioral vs. cognitive-behavioral, −.16, cognitive-behavioral vs. general verbal, .09, cognitive vs. behavioral, .22, and dynamic vs. cognitive, .08). This mean of .14 is not significantly different from means of the other comparisons listed, so such a finding is in synchrony with the other findings in the study.

**MAIN EXPLANATIONS FOR THE SMALL EFFECT SIZES FOR DIFFERENCES IN OUTCOMES OF ACTIVE TREATMENTS**

The effect sizes for comparisons of active treatments, both corrected and uncorrected, for the 17 meta-analyses, were usually relatively small and nonsignificant. The adjective “small” is preceded by the ambiguous qualifier “relatively” because the choice of a corresponding effect-size level varies among authors. Cohen (1977), for example, would call a d of .20 small (equivalent to a Pearson’s r of .10) but Rosenthal (1990, 1995) would call it greater than small because the designation is somewhat dependent on the requirements of the situation; for example, if only 4 of 100 persons having a heart attack are saved by taking aspirin, that is not a small percentage if you are one of the 4 people. Below we consider some probable explanations for this relatively small and nonsignificant relationship.

**Explanation 1: The Types of Treatments Do Not Differ Much in Their Main Effective Ingredients, and Therefore Small Differences with Nonsignificant Effects Are the Rule.** The common components between the treatments compared may be the most influential basis for explaining the small and nonsignificant difference effect. This was the explanation offered by Rosenzweig (1936) and later restated by Frank and Frank (1991), Luborsky et al. (1975), Strupp and Hadley (1979), and Lambert and Bergin (1994). Lambert and Bergin especially stressed the role of common factors across different psychotherapies in explaining the trend toward non-significant differences among the outcomes of different forms of psychotherapy. Elkin et al. (1989) and Imber et al. (1990) also considered the common factors across interpersonal and cognitive-behavioral psychotherapy in their explanations for the nonsignificant differences between different treatments in the National Institute of Mental Health Treatment for Depression Collaborative Research Program. This explanation emphasizes that the common components of different treatments may be so large and so much more potent than specific ingredients that the comparisons result in small and nonsignificant differences. Other components have also been suggested as common across treatments: the helping relationship with the therapist, the opportunity to express one’s thoughts (sometimes called abreaction), and the gains in self-understanding.

**Explanation 2: The Researcher’s Allegiance to Each Type of Treatment Compared Differs, Sometimes Favoring One Treatment and Sometimes Favoring the Other.** The researcher’s allegiance to each of the treatments in comparative treatment studies appears to influence the small effect sizes of each treatment outcome in the expected direction, as shown in the comprehensive evaluation by Luborsky, Diguier, Seligman, et al. (1999). To explain this more concretely: Treatment A in a meta-analysis may be favored by the researcher’s positive allegiance in one study, while in another study treatment A may suffer from a researcher’s negative allegiance.

**Explanation 3: Clinical and Procedural Difficulties in Comparative Treatment Studies May Contribute to the Nonsignificant Differences Trends.** There have been a series of rebuttals trying to explain the methodological problems that lead to the Dodo bird trend (e.g., Beutler, 1991; Elliott, Stiles, & Shapiro, 1993; Norcross, 1995; Shadish & Sweeney, 1991). These discussions tend to agree that although research shows that the small and nonsignificant difference effect exists, the effects of different treatments may appear in ways that have not yet been studied. Kazdin (1986), Kazdin and Bass (1989), Wampold (1997), and Howard et al. (1997) further explain that nonsignificant differences between treatments may reflect procedural and design limitations in comparative treatment outcome studies. These limitations include the representativeness of the measures of treatment process and outcome and the statistical power of the findings. Howard et al. (1997) further suggest doing separate meta-analyses for each contrasting pair of types of treatments, such as we have done for cognitive and cognitive-behavioral versus dynamic and other treatments.
Explanations: Interactions between certain patient qualities and treatment types, if not taken into account, may contribute to the nonsignificant difference effects. Several studies, such as those by Beutler et al. (1991), Blatt (1992), Blatt and Felsen (1993), and Blatt and Ford (1994), have shown that the match of the patient’s personality with different treatments can succeed in producing significant effects; when such matches are not taken into account, they may contribute to the nonsignificant difference effects.4

The Status of the Empirically Validated Treatment Movement

Much of the most recent comparative treatment research has been done as part of the increasingly fashionable empirically validated treatment movement (Luborsky, in press). One list of such studies in Chambless et al. (1996) might be thought by some to belong in our review, but it actually belongs in a separate category because the Chambless et al. study was not supposed to be a meta-analysis of the relative effectiveness of different active treatments. We mention this review just because it is commonly mistaken to report the relative efficacy of different treatments. However, it does not; the task force itself clearly stated that its focus was only on compiling a list of treatments that had been “empirically validated.”

Conclusions and Discussion

The available evidence has been summarized here from 17 meta-analyses of comparisons of active treatments with each other, in contrast to the more usual comparisons of active treatments with control treatments. The studies reviewed mainly included patients with the common diagnoses of depression, anxiety disorders (including obsessive-compulsive and phobic disorders), and mixed neurosis, but not patients who are psychotic or children. Also, the sample of studies included only those where patients were treated by these usual treatments: behavior therapy, cognitive therapy, cognitive-behavior therapy, dynamic therapy, rational-emotive therapy, and drug therapy.

Comparisons of active treatments with each other tend to have “small” and non-significant differences. For our sample of 17 of such meta-analyses in 6 reports of meta-analyses of comparisons of active treatments with each other, there is a mean uncorrected absolute effect size of .20 by Cohen’s d (Table 1). This is impressive because of its smallness as well as the fact that the six reports include meta-analyses with many studies. Another large-scale review of studies of treatment comparisons of active treatments also found a similar level of effect sizes: .19 by a Pearson r (Wampold et al., 1997). The mean effect size in our review supports our impression that a majority of comparisons of an active treatment versus an active treatment have relatively small effect sizes and nonsignificant differences between different psychotherapies, especially after corrections for the researcher’s allegiances, thus reaffirming the original Dodo bird verdict.

We also calculated medians of effect sizes (Table 1) to show that no one meta-analysis method skewed our overall mean effect size. Looking at Table 1, we see that the mean and median effect sizes, both weighted by the size of the sample of each study and unweighted, are almost identical for both the corrected and uncorrected effect sizes. Thus, our sample of meta-analyses has a good distribution, with no one meta-analysis method unduly affecting the overall mean uncorrected Cohen’s d effect size of .20.

To describe further what the overall uncorrected mean effect size actually represents, we must explain that it is a very conservative estimate. When calculating our mean effect size, we took the absolute value of each of the 17 effect sizes before summing them. This inflates our mean effect size because, if we had kept the signs and summed in that manner, certain effects would cancel each other, resulting in a lower mean effect size. Even then, our mean Cohen’s d of .20 is equivalent to a Pearson r of only .10. By Rosenthal and Rubin’s (1991) binomial effect size display (BESD) method, an r of .10 means that on average there is a 10% difference in success rate between psychotherapies (e.g., a change from 45% to 55%). Though a Cohen’s d of .20 may not be small according to Rosenthal (1990, 1995), Cohen (1977) does see it as small, and this average 10% difference in success rate is the most conservative estimate of the overall mean effect size due to our absolute values method of combining the Cohen’s d values.

Our general conclusion, therefore, is that Rosenzweig’s clinically based hypothesis of 1936 has held up. The outcomes of quantitative comparisons of different active treatments with each other, because of their similar major components, are likely to show mostly small and nonsignificant differences from each other.

Comparisons of active treatments with each other often need a correction. The reexamination of 29 mostly newer studies by Luborsky, Diguier, Seligman, et al. (1999) showed that a correction to the effect sizes is typically
needed because researcher’s allegiance to each of the therapies compared is highly correlated with treatment outcomes: the correlation was a Pearson’s r of .85! Researcher’s allegiance is therefore a reasonable basis for correcting effect sizes. After corrections for researchers’ allegiance were applied, the effect sizes were usually reduced and nonsignificant.

A few of the comparisons of active treatments with each other have larger and more significant differences. When considered one by one, a few of the correlations are moderate and reach the conventional level for significance. Such correlations are infrequent as part of the entire set of meta-analyses, but the presence of occasional significant differences in treatment outcomes perhaps should be taken seriously, as Lambert and Bergin (1994) tentatively suggest. Looking at the array of results for the meta-analyses that are surveyed in Table 1, one is struck by the variability of the effect sizes in which a few of them rise above the designation of a small and non-significant level to at least a moderate size. Wampold, Mondin, Moody, Stich, et al. (1997) noticed the same variability in their results but tended to view these as chance results in their large distribution of results. Crits-Christoph (1997) considered these variations more seriously, just as we are inclined to do, and suggested that these exceptions may reflect more than chance. Wampold et al. list 14 such exceptions in the 114 studies (p. 218). These suggest that something more than a Dodo bird verdict may be operating. Among the 114 studies in Wampold et al., Crits-Christoph (1997) identified only 29 studies with a noncollege student sample that involved comparison of noncognitive-behavioral treatments with each other (e.g., treatments other than cognitive therapy, desensitization, exposure, relaxation, skills training, and assertion training). Of these 29 studies, only 14 showed some significant difference between the treatment conditions, suggesting that the Dodo bird verdict may not apply as well in all cases.

The basic issue in this discussion is whether a few differences that were more than small and better than nonsignificant should be (a) attributed to chance factors or (b) pursued as illustrations of more than chance effects. There are arguments in favor of each alternative.

Comparisons of active treatment with controls appear to be less valuable for our main aim. “Controls” were used here to refer to (a) comparison groups that were purposely lacking in a component and (b) nonpsychotherapy treatments such as clinical management or wait-for-treatment groups. The type of comparative treatment study that is based on an active treatment versus a control naturally tends to give higher effect sizes (Grissom, 1996), so that the results from this type of effect size measure cannot justifiably be combined with the results of comparisons of active treatments. It is for this reason that we did not include the many studies using mainly treatment vs. control comparisons, such as Shapiro and Shapiro (1982), Engels, Garnefski, and Dickstra (1993), Van Balkom et al. (1994), Feske and Chambless (1995), and Grawe, Donati, and Bernauer (1994). Furthermore, the treatment-versus-control type of comparison tends to be not as revealing of the relative potency of a treatment as are comparisons of active treatments with each other.

Other important questions remain to be examined. The meta-analyses comparing different active treatments with each other suggest further research. The studies of comparative treatments still may not be sufficiently representative of the common diagnoses and the common types of psychotherapies. They may, for example, suffer from an unrepresentative selection of cases, as in Wampold, Mondin, Moody, Stich, et al. (1997). According to Crits-Christoph (1997), Wampold et al. have about half of their 114 studies involving the treatment of various forms of anxiety but too little of severe degrees of the usual diagnoses. Also, the studies in the sample may overrepresent behavioral and cognitive-behavioral treatments and underrepresent dynamic treatments. Therefore, more clinical trials are needed that correct for these distortions (Crits-Christoph, 1997). Our sample of studies partially corrects for such limitations. It is, impressive, however, that our sample of studies, which overlaps only in part with Wampold et al., also shows the small and nonsignificant difference effect that we call the Dodo bird verdict.

It will be valuable to have the research quality of each study that is included judged by independent judges whose therapy allegiances to each form of psychotherapy are known. So far, there is basically no correlation between the quality of the research study and the size of the outcomes in psychotherapy (Smith and Glass, 1977). We also found the same lack of correlation when we had two judges rate each study for the quality of the research on 12 dimensions and correlated their mean score with the outcome of the treatment (Luborsky, Diguer, Seligman, et al., 1999).

More investigations are needed of the degree of fit of each of the possible explanations of the small and non-
significant difference trend. For example, a comparison is needed for the degree to which the small and nonsignificant differences trend is best explained by common elements between the two treatments or by other factors as suggested by Crits-Christoph (1997).

We may ultimately find that research on the match of the type of patient to the type of treatment will offer more information than the usual comparative treatment research design with its focus on the comparison of different treatment types across patient types. After conducting more of these studies we may find that such match designs reveal more effective treatments for certain kinds of patients than the usual focus (Beutler, 1991; Blatt, 1992). Barber and Muenz (1996) found that in the TDCRP data (Elkin et al., 1989), although cognitive therapy and interpersonal patient therapy yielded similar results in that study, if one looks at the patients who are more obsessive, interpersonal patient therapy is better than cognitive therapy, and if one looks at the patients who are more avoidant, then cognitive therapy is better than interpersonal patient therapy. In other words, subgroups of patients might do better or worse with a specific treatment so that what is needed are hypotheses about those subgroups of patients.

Similar meta-analyses should be done with long-term treatments. We may find that long-term treatments have a long-term buildup of improvement, that may ultimately lead to more benefit (Luborsky, in press). Finally, it may be worth examining symptom outcomes along with other kinds of outcomes of psychotherapy. Small and nonsignificant outcomes do not mean that the treatments compared have the same effects on all patients. Specific outcome measures such as depression and anxiety may tempt us to forget that there may be other differences between treatments. Two patients, for example, after participating in different treatments, may feel better and not currently depressed, but one of them also may have made other important gains. One currently nondepressed man said that he achieved a better understanding of his relationship with his wife and was able to make helpful changes. We and others therefore should score other aspects of the patients’ changes as well, even beyond the symptom measures.

NOTES
1. We have used the usual wording “nonsignificant difference” rather than “equivalent” because the former is usually more fitting. But there are times when it is possible to imply an equivalence between two compared groups by following the method suggested by Rogers, Howard, and Vessey (1993). That suggested method reveals whether two groups are sufficiently similar to each other to be thought of as equivalent.

2. Consider a study comparing $T_1$ versus control, that finds $d = .80$ and a study comparing $T_2$ versus control, that finds $d = .30$. We conclude $T_1$ is better than $T_2$ because a $d$ of .80 is larger than a $d$ of .30. However, if the study of $T_2$ had used a much sicker population of patients, the smaller $d$ is not due to a difference between treatments but to a difference between clients. A head-to-head comparison of $T_1$ versus $T_2$ for a sample of patients for which both $T_1$ and $T_2$ would be appropriate might find no difference at all.

3. To provide a more exact calibration of the largeness–smallness of comparison of active treatment versus active treatment, these methods can be used: (a) The measure can be compared with the overall effect of the treatment (e.g., $d = .85$) (b) A second method uses the coefficient of robustness (mean $d/S$), an index of the clarity of the directionality of the results in relation to their homogeneity (Rosenthal, 1991).

4. The use of the term “efficacious” will remind many readers of the becoming-popular distinction between “efficacy” and “effectiveness”; this is essentially the supposed distinction between a research–context comparison of treatments (efficacy) with a clinic–context comparison of treatments (effectiveness). In the last 6 or 7 years especially, the opinion has spread that clinic-based treatment tends to be less effective than research-based treatment. The idea became even more prevalent after Weiss, Weiss, and Donenberg (1992), investigating the differential effectiveness of child psychotherapy under the two conditions, reported that “clinic therapy” was far less effective than “research therapy.” But, on the contrary, a much larger review (Shadish, 1996) showed that “clinic therapy” performed reasonably well compared with “research therapy” and the same conclusion was reported in Shadish et al. (1997).

ACKNOWLEDGMENT
This work was partially supported by the National Institute on Drug Abuse, Research Scientist Award 2K05 DA 00168-23A (to Lester Luborsky), and NIMH Clinical Research Center Grant MH45178 (to Paul Crits-Christoph). We acknowledge helpful suggestions from Irene Elkin. This article was presented at a celebration on May 26, 2000 in honor of Saul Rosenzweig’s many contributions, at Washington University’s Department of Psychology, St. Louis, Missouri.

REFERENCES


Svarberg, M., & Stiles, T. C. (1991). Comparative effects of...


Received August 22, 2000; revised January 17, 2001; accepted January 30, 2001.