Journal of Consulting and Clinical Psychology
1995, Vol. 63, No. 2, 317–320

Relative Effectiveness of Behavioral Versus Nonbehavioral Child Psychotherapy

Bahr Weiss
Vanderbilt University

John R. Weisz
University of California, Los Angeles

Some researchers have concluded that, for children and adolescents, behavioral interventions may be more effective than nonbehavioral interventions. Other researchers, however, have proposed artifactual hypotheses for the apparent superiority of behavioral treatments. In this study, one such hypothesis was evaluated: that the apparent superiority of behavioral interventions among children is due to differences in the methodological quality of studies of behavioral and nonbehavioral treatments. Meta-analytic results reported in this article found little support for this hypothesis.

When Lewis Carroll's dodo bird stated that "everybody has won, and all must have prizes" (Carroll, 1865/1988), it is doubtful that Carroll anticipated that his bird would become the most widely quoted dodo in the psychology literature. Beginning with Rosenzweig (1936), a number of authors have used this quotation to illustrate the hypothesis that all forms of psychotherapy are roughly equivalent in effectiveness (although Carroll's metaphor might also be taken as a commentary on a lack of coordination among therapy researchers; Stiles, Shapiro, & Elliott, 1986). Since that time, the "dodo verdict" has had supporters as well as critics (e.g., Bergin & Lambert, 1978). Perhaps the strongest support came from Smith, Glass, and Miller (1980), who concluded that different types of treatment do not produce substantially different degrees of benefit.

The Smith et al. (1980) psychotherapy meta-analysis included participants of all ages. The two meta-analyses that have examined the relation between type of treatment and outcome among children (herein we use the word "children" to include adolescents as well as children) found, in contrast, that behavioral interventions were associated with significantly larger effect sizes than nonbehavioral interventions (Casey & Berman, 1985; Weisz, Weiss, Alicke, & Klotz, 1987).[1] This suggests that, for children, behavioral treatments may be more effective than nonbehavioral treatments and that, indeed, not all participants have won the race. However, a number of artifactual explanations for this apparent superiority have been proposed.

Shirk and Russell (1992), for instance, hypothesized that these results may largely be due to differences in the methodological quality of behavioral and nonbehavioral outcome studies. Specifically, Shirk and Russell (1992) suggested (a) that nonbehavioral studies may be less rigorous methodologically than behavioral studies, (b) that methodological rigor is positively related to the magnitude of a study's effect size (or sizes), and (c) that this relation, rather than differences in the effectiveness of behavioral and nonbehavioral treatments, is largely responsible for behavioral studies being associated with larger effect sizes more than nonbehavioral treatments. They analyzed a set of 24 outcome studies that evaluated 29 nonbehavioral treatment groups. Similar to some, but not all, previous investigations (e.g., Bergin & Lambert, 1978; Weiss & Weisz, 1990), Shirk and Russell (1992) found that their rating of methodological quality was positively correlated with effect size.

These findings provide some support for one component of Shirk and Russell's hypothesis. However, two key elements of their hypothesis were left unanswered: (a) whether behavioral and nonbehavioral interventions differ in methodological quality, and (b) whether such differences account for effect size differences between behavioral and nonbehavioral studies. In the present study, we evaluated these two components to test this artifactual hypothesis more fully.

Method

The sample of studies ($n = 105$) on which the present analyses are based was taken from our meta-analysis data set (Weisz et al., 1987);

Bahr Weiss, Department of Psychology and Human Development, Vanderbilt University; John R. Weisz, Department of Psychology, University of California, Los Angeles.

This research was supported in part by National Institute of Mental Health Grants R18-MH50265 and R01-MH48522.

We thank Karin Süsser and Julie Tapp for their assistance in coding the methodological variables, Aviva Wasserman for her comments on various drafts of the manuscript, and Stephen Shirk for providing information on the procedures used in the S. R. Shirk and R. L. Russell (1992) study.

Correspondence concerning this article should be addressed to Bahr Weiss, Department of Psychology and Human Development, Box 512 GPC, Vanderbilt University, Nashville, Tennessee 37203. Electronic mail, including requests for an extended report, may be sent via Bitnet to WEISSBH@VUCTRVAX.

[1] Some researchers (e.g., Kazdin, 1993) have questioned the meaningfulness of comparisons of broad classes of treatments such as behavioral versus nonbehavioral therapies, because the different treatments within such broad classes may be quite different from each other. However, in our meta-analysis (Weisz et al., 1987), there were significant differences between these broad classes but not among the narrower forms of treatment which constituted the behavioral and nonbehavioral groups. This suggests that, although certain aspects of the treatments within the broad groups may vary, important aspects that influence outcome may be similar enough to make such comparisons meaningful.

the Shirk and Russell (1992) sample was a subset of this sample. For a description of our sample and coding for type of treatment, see Weisz et al. (1987). For the analyses described later, we derived two overlapping sets of methodological factors (exact coding definitions of these variables are available in the extended report for this study). For each of these methodological variables, two judges independently coded 15% of the studies in our sample; kappas were in the good-to-excellent range (Fleiss, 1981).

Coding for Shirk–Russell Factors

Our first set of methodological factors was that described by Shirk and Russell (1992): (a) lack of random assignment (whether participants were or were not randomly assigned; kappa = 0.82), (b) rater evaluation bias (whether the outcome rater was aware of the participants' experimental group assignment or of the experimental hypothesis; kappa = 0.78), (c) uncontrolled concurrent treatment (treatment that participants received [e.g., from other therapists] which was not controlled by the experimenter; kappa = 0.83), (d) unequal attrition for treatment and control groups (whether attrition for the treatment and control groups differed by more than 10%; kappa = 1.00), (e) therapist inexperience (whether the therapist was relatively inexperienced—i.e., had no formal training [e.g., a parent]—or was a graduate student; kappa = 0.83), (f) mono-operationalization (use of a single outcome measure; kappa = 1.00), (g) insufficient treatment (less than 10 sessions; kappa = 1.00), and (h) failure to ensure treatment integrity (lack of a treatment manual or audiotaped supervision; kappa = 0.81). Ratings for these factors were weighted and summed into an overall score with the method described by Shirk and Russell (1992).

Coding for Weiss–Weisz Factors

For reasons detailed later, we felt that this operationalization of methodological quality might not provide a maximally powerful test. Consequently, we also evaluated Shirk and Russell's hypothesis using a second set of methodological factors, which included a subset of their factors as well as others from our previous meta-analyses (Weiss & Weisz, 1990; Weisz et al., 1987). This second set included (a) therapist inexperience (coded the same as the Shirk–Russell factor), (b) rater blindness (coded the same as Shirk & Russell's "rater evaluation bias" factor), (c) participant blindness to outcome assessment (whether participants were aware that the outcome assessment was being made, which could result in biased behavior), (d) failure to ensure treatment integrity (coded similarly to Shirk & Russell's factor, except that we also coded that treatment integrity had been ensured if an intervention was presented by means of an audiotape or videotape [e.g., as in videotaped modeling]), (e) participant assignment (coded along a scale ranging from 1 to 3 with regard to the manner in which participants were assigned to treatment and control groups: 1 = nonrandomly, 2 = randomly without matching of treatment and control groups, and 3 = randomly with pretreatment matching of groups on at least one dependent variable), (f) participant attrition (the average percentage of treatment and control-group participants who failed to complete the posttreatment assessment, which we felt was a greater threat to validity than differential attrition, which Shirk and Russell (1992) coded. A study with 90% attrition for both treatment and control groups would be of questionable validity yet have zero differential attrition. However, we also conducted the analyses described later using differential attrition; results were virtually identical to those obtained with overall attrition), and (g) measurement technology (the assessment methodology used to make the outcome assessment, coded on a soft-to-hard scale ranging from 1 [*self-reports*] to 4 [*objective life-event data, such as arrests*]; Shapiro & Shapiro [1982]). To derive an overall rating of methodological quality, we standardized these seven factors and then took the mean of the seven. However, we also performed analyses based on the individual factors, which we felt might provide a more powerful test of the hypothesis.

There were several factors which Shirk and Russell (1992) included in their assessment of methodological quality that we did not include in our second set. These factors were (a) *mono-operationalization*, which we did not include because there appeared to be no compelling reason to conceptualize this as a validity factor; (b) uncontrolled concurrent treatment, which we did not include because treatment and control groups should within random variability be equally exposed to this factor; and (c) insufficient treatment, because the appropriate number of sessions would vary with treatment objectives and type of treatment (e.g., one would expect that behavioral treatments generally would require fewer sessions than nonbehavioral treatments).

Computation of Effect Sizes

We computed effect size estimates for each dependent variable by dividing the mean posttherapy treatment group/control group difference by the standard deviation of the control group. Some researchers (e.g., Casey & Berman, 1985) favor the use of the pooled treatment and control standard deviation as the denominator for the effect size. However, if as some evidence suggests (e.g., Weiss & Weisz, 1990), one consequence of therapy is an increase in behavioral variability, such pooling can cause interpretational and statistical problems (see Smith et al., 1980), which we sought to avoid. When means and standard deviations were not available in the published report or unobtainable from the author (or authors), we used the formulas provided by Smith et al. (1980) for computing effect sizes from *t* statistics and so forth. We applied Hedges and Olkin's (1985) adjustment for small sample bias with our effect sizes.

Results

Overview of Analyses

The first set of analyses that we report (discussed later) was based on Shirk and Russell's (1992) operationalization of methodological quality and on their codings for the nonbehavioral treatments, which differed from ours in several instances.[2] Paralleling their approach to dealing with multiple outcome measures in a single study, we structured our analyses so that each treatment group served as a single observation. Thus, methodological quality and effect size were averaged across different outcome measures for each treatment group. The second set of analyses was based on our seven methodological factors, as described earlier. In these analyses, we used two different approaches for dealing with multiple outcome measures from a single study. First, we used the same strategy as Shirk and Russell, collapsing up to the level of the treatment group. Second, we collapsed up to the outcome measure level. Thus, in this second set of analyses, each treatment group provided as many

[2] In discussing their coding for type of treatment, Shirk and Russell (1992) stated that "Weisz et al. (1987) identified 27 nonbehavioral treatment groups" and that "Reexamination of the 108 [Weisz et al., 1987] studies by two raters yielded 29 nonbehavioral treatments" (p. 704). Actually, as we reported previously (Weisz et al., 1987, p. 544, Table 1), we coded 28 treatment groups as nonbehavioral. Comparison of our nonbehavioral treatment groups to Shirk and Russell's (1992) groups indicated that there were five discrepancies between our sample of nonbehavioral studies and theirs. Details of the discrepancies are available from Bahr Weiss.

observations as it contained dependent measures. All analyses based on our set of factors were conducted twice, using both of the approaches for collapsing just described. Results of the two approaches were highly similar, with the second approach (i.e., collapsing up to the outcome measure level) in each instance producing effects that were equally or more significant than the results produced by the first approach. In this article, we report only the effects of the first approach.

Results for Shirk–Russell Factors

To evaluate the Shirk–Russell hypothesis more fully, we first conducted a one-way ANOVA to determine whether the behavioral and nonbehavioral treatment studies differed in regard to overall level of methodological quality, as defined by Shirk and Russell (1992). The two groups did differ significantly, $F(1, 153) = 9.67$, $p < .005$. However, contrary to the Shirk-Russell arguments, the mean rating for methodological quality for the nonbehavioral treatments was higher than the mean for the behavioral treatments. Analysis of the individual Shirk and Russell (1992) factors indicated that this difference was primarily due to the "insufficient treatment" variable, with behavioral treatments more frequently falling into the "insufficient treatment" category, $\chi^2(1, N = 155) = 18.74$, $p < .0001$. Because, as noted earlier, it could be argued that fewer than 10 sessions might be "sufficient" treatment for behavioral interventions and, thus, not an appropriate criterion for methodological quality for these treatments, we dropped this variable from the overall measure of methodological quality. Using this revised measure, we found that the two groups did not differ significantly.

Next, we tested whether behavioral and nonbehavioral treatments differed in regard to effect size, after controlling for methodological quality. First, however, we tested the effect of type of treatment ignoring the effect of methodological quality. We found that behavioral treatments were associated with significantly larger effect sizes than nonbehavioral treatments (mean effect sizes, .85 and .44 for behavioral and nonbehavioral treatments, respectively), $F(1, 153) = 4.90$, $p < .05$. When we controlled for Shirk and Russell's measure of methodological quality, the difference between the two types of treatment was marginally significant, $F(1, 152) = 3.69$, $p < .06$, with behavioral treatments associated with larger effects than nonbehavioral treatments (adjusted effect sizes, .84 and .47, respectively). When we reconducted this analysis using the revised measures of methodological quality (i.e., dropping the "insufficient treatment" factor), the difference between the groups was significant, $F(1, 152) = 4.80$, $p < .05$.

We also entered the eight Shirk–Russell methodological factors individually into these models rather than their weighted sum; results were similar to those reported earlier. In summary, when we controlled for the Shirk–Russell methodological factors, behavioral treatments continued to be associated with a larger mean effect size than nonbehavioral treatments.

Results for Weiss–Weisz Factors

We conducted analyses on the basis of what we considered a more appropriate set of methodological factors. We first conducted a one-way ANOVA to determine whether the behavioral and nonbehavioral treatment studies in our sample differed in regard to their overall level of methodological quality. In this analysis, type of treatment served as the independent variable and our overall rating of methodological quality as the dependent variable. The difference between the two groups was marginally significant, $F(1, 152) = 3.29$, $p < .08$, with the mean rating for methodological quality for the nonbehavioral treatments higher than the mean for the behavioral treatments. Because this analysis was based on the mean of our seven methodological factors, group differences on the individual factors may have been obscured. Consequently, we conducted a discriminant function analysis (Harris, 1985) using the seven individual methodological factors. The groups were significantly discriminated by the single canonical variate, $F(7, 71) = 2.60$, $p < .05$. Inspection of the canonical structure (Harris, 1985) indicated that the canonical variate was composed primarily of participant blindness to outcome assessment and therapist experience (i.e., both of these variables were positively correlated with the canonical variate), with the nonbehavioral studies associated with higher levels of the canonical variate (i.e., with higher levels of methodological quality).

We next determined whether behavioral treatments were associated with larger effect sizes than nonbehavioral treatments, once the effect of methodological differences was controlled. First, we assessed whether the two forms of treatment differed in regard to effect size before methods differences were controlled. Behavioral treatments were associated with larger effect sizes than nonbehavioral treatments, $F(1, 152) = 5.13$, $p < .05$; effect sizes, .85 and .42 for behavioral and nonbehavioral treatments, respectively). When we controlled for the effect of the methodological factors by including our measure of overall methodological quality in the model, the difference between the two types of treatment remained significant, $F(1, 151) = 6.39$, $p < .05$; adjusted effect sizes, .86 and .38, respectively. The two types of treatment also differed significantly in regard to effect size after controlling for the canonical variate derived from our discriminant function analysis, $F(1, 76) = 5.72$, $p < .05$; adjusted effect sizes, .70 and .34, respectively.

We also analyzed our data by restricting our sample to only those studies that directly compared behavioral and nonbehavioral treatments. By doing so, each comparison was matched for sample and site characteristics, and so forth. (see Shadish & Sweeny [1991] and Shapiro & Shapiro [1982] for a more detailed rationale for this strategy). Although based on a much smaller sample ($n = 10$), our substantive findings were the same. In this restricted sample, nonbehavioral treatments had higher scores on our overall methodological quality variable than behavioral treatments, although this difference was nonsignificant. The mean effect size for behavioral treatments was significantly larger than that for nonbehavioral treatments (.76 and .17, respectively), $F(1, 18) = 5.57$, $p < .05$. This difference remained significant when the effect of methodological quality was controlled, $F(1, 17) = 5.77$, $p < .05$.

Discussion

Our results suggest that, at least in our sample of studies, using the present definitions of methodological quality, the apparent superiority of behavioral treatments in children is not an artifact of methodological quality. However, because meta-analysis is a correlational technique, our results should be considered suggestive rather than definitive; it is possible that some untested confounding variable is responsible for behavioral treatments being associated with larger effect sizes. For instance, a number of authors (e.g., Robinson, Berman, & Neimeyer, 1990) have suggested that the apparent superiority of certain forms of treatment may be due to investigator allegiance effects; that is, when one treatment is compared with another, a researcher's preference for or differential expertise with one particular intervention may influence the apparent relative effectiveness of the treatments. However, although a number of studies have reported a relation between allegiance and outcome, the direction of causality between these two constructs is not clear. Investigators' allegiances may correlate with outcome because their expectations have been influenced by reports of effective treatments in the literature and by the results of their own past studies. To fairly control for allegiance effects in future research, it may be useful for investigators to directly compare, in the same study, behavioral and nonbehavioral interventions supervised and conducted by therapists who have explicit allegiance to the type of intervention they conduct.

It also is important to note that nonbehavioral treatments have not received as extensive an evaluation as behavioral treatments and that both forms of treatment need more evaluation under the conditions that prevail under clinical practice (Weisz & Weiss, 1993; Weisz, Weiss, & Donenberg, 1992). In addition, it will be useful to determine whether the statistically significant difference between behavioral and nonbehavioral treatments is clinically significant (i.e., whether the difference between behavioral and nonbehavioral treatments is meaningful in terms of client functioning; Kendall & Maruyama, 1985). Finally, if future tests continue to indicate that behavioral methods are more effective than nonbehavioral methods, it will be important to assess which features of behavioral treatment (e.g., targeting of specific problems, use of explicit reinforcers, and direct teaching of coping skills) constitute the active ingredients accounting for the superior effects. What the present data do suggest is that, contrary to what other investigators have proposed, the fact that behavioral treatments are associated with larger effect sizes more than nonbehavioral treatments cannot be readily explained by differences in methodological quality.

References

Bergin, A. E., & Lambert, M. J. (1978). The evaluation of therapeutic outcomes. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis.* (pp. 139–190) New York: Wiley.

Carroll, L. (1988). *Alice's adventures in Wonderland.* New York: H. N. Abrams. (Original work published in 1865.)

Casey, R. J., & Berman, J. S. (1985). The outcome of psychotherapy with children. *Psychological Bulletin, 98,* 388–400.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Harris, R. J. (1985). *A primer of multivariate statistics* (2nd ed.). New York: Academic Press.

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

Kendall, P. C., & Maruyama, G. (1985). Meta-analysis: On the road to synthesis of knowledge? *Clinical Psychology Review, 5,* 79–89.

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

Rosenzweig, S. (1936). Some implicit common factors in diverse methods of psychotherapy. *American Journal of Orthopsychiatry, 6,* 412–415.

Shadish, W. R., & Sweeny, R. B. (1991). Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology, 59,* 883–893.

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

Shirk, S. R., & Russell, R. L. (1992). A reevaluation of estimates of child therapy effectiveness. *Journal American Academy of Child and Adolescent Psychiatry, 31,* 703–709.

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

Stiles, W. B., Shapiro, D. A., & Elliott, R. (1986). "Are all psychotherapies equivalent?" *American Psychologist, 41,* 165–180.

Weiss, B., & Weisz, J. R. (1990). The impact of methodological factors on child psychotherapy outcome research: A meta-analysis for researchers. *Journal of Abnormal Child Psychology, 18,* 639–670.

Weisz, J. R., & Weiss, B. (1993). *Child psychotherapy: What we know and what we need to know.* New York: Sage.

Weisz, J. R., Weiss, B., Alicke, M. D., & Klotz, M. L. (1987). Effectiveness of psychotherapy with children and adolescents: A meta-analysis for clinicians. *Journal of Consulting and Clinical Psychology, 55,* 542–549.

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist, 47,* 1578–1585.

Received August 6, 1993
Revision received February 11, 1994
Accepted February 23, 1994 ■