Journal of Abnormal Child Psychology, Vol. 18, No. 6, 1990, pp. 639-670

The Impact of Methodological Factors on Child Psychotherapy Outcome Research: A Meta-Analysis for Researchers

Bahr Weiss[1,3] and John R. Weisz[2]

Two recent meta-analyses have generated evidence for child and adolescent psychotherapy effects. However, critics note that such meta-analyses often include studies with methodological shortcomings which might invalidate their results. In the present study, we explored whether the results of the most extensive child/adolescent meta-analysis might have been influenced by such methodological variables, focusing on internal validity and external validity factors. Together, these factors accounted for two-thirds as much variance as the substantive factors (e.g., type of therapy, age) in the original meta-analysis. This suggests that relative to these therapy and child-characteristic variables, methodological factors have a substantial, though smaller, impact on meta-analysis results. In general, increased experimental rigor was related to larger *effect sizes; this argues* against *the hypothesis that methodologically weak studies have led to an overestimate of therapy effects. No significant interactive relations were found between validity factors and predictors of outcome; this suggests that the relations noted in previous meta-analyses between outcome and various variables were* not *distorted by the validity factors tested here.*

Manuscript received in final form May 14, 1990.
The research described here was supported in part by grants from the North Carolina Department of Human Resources, Division of Mental Health, Mental Retardation, and Substance Abuse Services (41626), and from the National Institute of Mental Health (1 RO3 MH38450). The authors gratefully acknowledge the assistance of Todd Morton, who provided reliability coding.

[1]Department of Psychology and Human Development, Box GPC-86, Peabody College, Vanderbilt University, Nashville, Tennessee 37203.
[2]Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599.
[3]Correspondence regarding this article should be addressed to Bahr Weiss, Department of Psychology and Human Development, Box 86, Peabody College, Vanderbilt University, Nashville, Tennessee 37203.

0091-0627/90/1200-0639$06.00/0

Two recent meta-analyses support the effectiveness of psychotherapy with children and adolescents. Casey and Berman (1985) found an overall effect size (*ES*) of 0.71, indicating that across the various outcome measures, at the post-treatment assessment the average treated child in their sample was functioning better than 76% of the untreated children. Similarly, Weisz, Weiss, Alicke, & Klotz (1987a) found an effect size of 0.79; the average treated youngster in their sample was functioning better than 79% of those not treated.

Both of these analyses, as well as most meta-analyses of psychotherapy outcome (e.g., Shapiro & Shapiro, 1982), have chosen not to restrict their sample of studies to those that might be considered free of methodological flaws. Some authors, however, have expressed misgivings about the inclusion or equal weighting of methodologically weak studies in meta-analyses (see, e.g., Rachman & Wilson, 1980; Wilson & Rachman, 1983). Specifically, these authors are concerned that methodological factors may influence outcome, and that if methodologically weak studies are included, the findings of outcome meta-analyses will be in part a function of the design shortcomings in the literature reviewed. If this concern is valid, then the conclusions drawn from previous meta-analyses may need to be modified, or at least qualified.[4]

Some aspects of the effects of methodological factors on the results of outcome studies have been assessed previously (see, e.g., Shapiro, 1985; Smith, Glass, & Miller, 1980). However, none of these investigations focused specifically on child therapy studies; since adult and child populations (and treatments) may differ along a number of important dimensions, generalizability across developmental level is unclear. For instance, recent data (e.g., Gould, Shaffer, & Kaplan, 1985) suggest that in child therapy, attrition is more highly related to characteristics of the *parents* of the client than of the client him- or herself, a finding not likely to be found among adults. Thus, the influence of attrition on the results of child and adult outcome studies may not be similar.

There is an even more critical issue which remains unaddressed: In the past, when assessing the impact on treatment outcome of differences in va-

[4]One might wonder why we, or anyone, would bother expending much energy assessing the impact of methodological factors. Why, as Mintz (1983) aptly questioned, should one bother investigating the effects of impaired validity since, it might be argued, researchers really should pay little or no attention to studies containing design flaws? If their results contradict those of well-designed studies, then shouldn't the less well-designed studies be disregarded? If their results agree, what do the less well-designed studies add? However, we believe that in fact there are few studies free of design flaws or limitations: As other authors have noted (e.g., Shapiro and Shapiro, 1983), design compromises appear inevitable. Thus, one would be left with a small sample indeed if one restricted oneself to those studies deemed methodologically pure.

lidity between studies, meta-analytic investigators have generally compared effect sizes of studies with different validity characteristics (e.g., Shapiro & Shapiro, 1982; Smith et al., 1980; Weisz et al., 1987a). Such tests, however, may be somewhat limited in their implications. While similarity of effect sizes between groups differing on a factor of interest suggests that the factor does not influence outcome, such similarity is not by itself a sufficient test. Of at least equal importance is the question of whether the factor in question influences relations between variables (Mintz, 1983)—i.e., whether it interacts with other variables. In the present study, we address both questions.

Answers to this more complete set of questions should help determine whether the conclusions of previous meta-analyses may need to be altered. We also hoped that assessment of the effects of design shortcomings might assist researchers in planning outcome studies. As other authors have noted (e.g., Kraemer & Andrews, 1982; Shapiro and Shapiro, 1983), design compromises in psychotherapy outcome research appear inevitable. Understanding the effects of different methodological shortcomings may allow one to make informed choices among the various alternatives imposed by the constraints of one's research question. For example, teachers are an important source of information about a child's functioning, yet it is often difficult to keep a teacher unaware of whether a child has been assigned to a treatment or control group. How much of one's limited resources should be allocated to keeping a teacher experimentally blind should depend in part on the seriousness of the effects of teacher awareness of group membership.

Among the various methodological factors whose impact on outcome one might assess, those relating to internal and external validity (Campbell & Stanley, 1963) are probably among the most important. Internal validity, in the case of outcome research, refers to the extent to which changes in subjects' psychological or behavioral functioning can be attributed to the experimental treatment. In the present investigation, we examined the effect of several potentially important internal validity variables: subject attrition, random vs. nonrandom assignment of subjects to experimental groups, and three factors[5] related to reactivity: (1) the measurement technology of the

[5]In an analysis of components of reactivity, Shapiro and Shapiro (1982) found that the specificity of the outcome measure (in relation to the goals of treatment) appeared to be the aspect of reactivity most highly correlated with outcome. However, as we noted in our previous meta-analysis, such a similarity between treatment activities and goals, and outcome measures may in some cases be valid. Such an overlap would be appropriate, we maintained, when the outcome assessment represents the most valid assessment of the treatment. For instance, in evaluating the efficacy of an in vivo desensitization treatment of a dog phobia, one would likely assess the child's ability to approach a dog. This would involve a high degree of overlap between treatment activities and goals, and outcome assessment, yet we believe that this would actually be the most valid assessment of the child's fear of dogs. Consequently, we did not include this variable as a validity factor in our investigation.

outcome measure (e.g., self-reports vs. life-event data), likely related to the ease with which biasing could occur; (2) the experimental blindness of the raters; and (3) the experimental blindness of the subjects.

We chose the subject attrition and subject assignment dimensions because they are, as Smith et al. (1980) have remarked, among the most potentially serious threats to internal validity. Outcome measure reactivity was selected because it was the single largest correlate of effect size reported by Smith et al. (1980), and consequently seemed a potentially important methodological factor.

In the case of therapy outcome research, external validity refers to the extent to which the results of an experimental treatment can be generalized to actual clinical practice. The results of a recent study suggest that investigation of this area—external validity—may be critical: In this study (Weisz & Weiss, 1989a), we found that in contrast to the apparent effectiveness of child psychotherapy under research conditions (see Weisz et al., 1987a), therapy that occurs under naturalistic conditions may be no more effective than no treatment. Thus, investigation of the impact of external validity factors appears crucial.

In the present investigation, we considered three dimensions along which Kazdin (1978a) has suggested studies may be compared vis-à-vis generalizability: (1) whether the children serving as subjects would have been in treatment irrespective of the research project; (2) whether the individual administering the treatment was a practicing clinician; and (3) whether the treatment took place in a clinical setting. These specific dimensions were selected because they appeared to be the most overarching, in that many other potential dimensions seemed largely subsumed by them (e.g., "manner of recruitment" [Kazdin, 1978a] would likely be highly correlated with "treated anyway").

We also evaluated the impact of the type of control group to which the treatment group was compared. Although not strictly categorizable as either an internal or external validity factor, there is reason to at least suspect that this variable might influence outcome: Some control groups, such as attention-placebo or minimal treatment groups, are designed to contain potentially active elements which could affect subjects' behavior. If in fact these control groups do alter subjects' behavior, then the apparent effectiveness of therapy may depend in part upon the type of control group with which the treatment group is compared.

As Kazdin (1978b) has noted, however, calling a control group an "attention-placebo group" does not ensure that it is in fact controlling for the extra attention (or any other nonspecific factor) that therapy involves. By contrasting outcomes of treatment groups compared to different control groups, we hoped to determine whether type of control group does in fact have an impact; if it does not, this would suggest that these control groups (i.e., attention-placebo groups) are not controlling for nonspecific factors. Thus,

our control group analyses had two objectives: (1) to determine whether our previous meta-analytic results need to be qualified with respect to type of control group; and (2) to determine whether attention-placebo control groups do control, at least in part, for nonspecific factors.

And finally, we attempted to determine whether therapeutic interventions increase the variability of outcome measures, as has been suggested by some authors (e.g., Lambert, Shapiro, & Bergin, 1986). Such an increase, if found, could have important implications for how we interpret results of outcome studies, in at least two ways. First, if variability is in fact increased, then inspection of only the mean will fail to tell the full story of the effects of therapy. And second, determining whether therapy increases variability could help determine the most appropriate definition for meta-analytic effect sizes. Effect size is generally defined as the mean of the treatment group on some measure of outcome minus the mean of the control group, divided by the standard deviation. While some researchers (e.g., Hedges, 1982) favor the use of the pooled control and treatment group standard deviation, others (e.g., Smith et al., 1980) believe that the unpooled control group standard deviation is to be preferred. This is in part because pooling will be inappropriate if treatment and control group variance differ, which treatment group increases in variability would suggest might be the case.

METHOD

For the purposes of selecting studies for our analysis, we defined psychotherapy as any intervention designed to alleviate psychological distress, reduce maladaptive behavior or enhance adaptive behavior, through counseling, structured or unstructured interaction, a training program, or predetermined treatment plan. Only published studies using a control group, and with a mean subject age between 4 and 18 years were included. Our literature search, for January 1960 through September 1985, was based on (1) a computer search; (2) manual inspection of *Psychological Abstracts, Behavior Therapy, Journal of Abnormal Psycholoy, Journal of Consulting and Clinical Psychology,* and *Journal of Counseling Psychology*; and (3) review of the bibliographies of the Casey and Berman (1985) and Smith et al. (1980) meta-analyses, as well as the bibliographies of studies passing our inclusion criteria. Followup assessments to original reports were considered part of the original study; this produced a final pool of 105 studies[6] (see the appendix to this article).

[6]In our original article (Weisz et al., 1987a), we incorrectly reported that the number of studies in our sample was 108; the correct number is 105. Three studies which were ultimately excluded from our analyses were mistakenly retained in our bibliography.

Although there are several different approaches to computing and analyzing effect sizes (e.g., Cohen, 1977; Hunter, Schmidt, & Jackson, 1982; Smith et al., 1980), we used the procedures of Glass and Smith (Smith et al., 1980) for calculating and analyzing effect sizes.[7] The Glass-Smith approach developed out of psychotherapy research, and thus seems most appropriate for our study; further, our previous meta-analysis was based on the Glass approach, and we wanted the present findings to be interpretable in the light of our previous findings. In general, we used as our estimate of the effect size the post-treatment mean of the treatment group minus the post-treatment mean of the control group, divided by the control group standard deviation.

Because studies with larger samples may yield more precise estimates of treatment effects, we considered weighting each study by some function of sample size (e.g., $N^{1/2}$; $1-[1/N]^{1/2}$). However, when we reran the analyses described below using various weighting schemes, results changed very little; hence, we report only the results of the unweighted analyses.

Most studies in our pool included multiple-outcome measures. Consequently, in most cases, studies initially produced a number of effect sizes. To retain all effect sizes in our analysis would have resulted in disproportionate weighting of those studies with the most measures; as well, it could possibly have led to an underestimation of the error variance (Strube & Hartmann, 1983). Consequently, we collapsed across dependent measures, except when such collapsing would have been inappropriate; e.g., when analyzing a variable such as rater blindness, which was not necessarily constant across the different dependent variables within a particular study.

We felt that this approach provided an appropriate balance of concern for Type I and Type II errors, given the importance of avoiding a Type II error (i.e., failing to appropriately reject the null hypothesis that a validity factor did not influence outcome). It is important to note, however, that other analytic strategies have been sometimes employed to address the issue of multiple-effect sizes within a single study. Some researchers (e.g., Casey & Berman, 1985) have, for example, in certain instances selected one effect size from each study, discarding all other data points from the analyses. We chose not to use this data-dropping strategy for our primary analyses because, although it eliminates the problem of nonindependence, it does so at the expense of discarding a large amount of information.

Another potential strategy for addressing the issue of nonindependence of effect sizes involves collapsing up to *and within* the level of the analysis,

[7]Hunter et al. (1982) present a series of techniques designed to assess the impact of, and to correct the effect size and its variance for, sampling error, differences in reliability across studies, and other artifacts that may influence the effect size. However, as these authors note (Hunter et al., 1982, p. 31), their procedures require information (e.g., reliabilities) that generally is not included in published reports. We felt that estimation of such values might contribute more noise than it removed; consequently, we did not apply these techniques.

such that each study provides at most one effect size for each level of the variable being analyzed. If, for example, one were analyzing the source-of-outcome measure (e.g., parent vs. teacher vs. self-report vs. trained observer), all effect sizes (i.e., outcome measures) within a particular category (parent, teacher, self-report, or observer) would be collapsed within each study; thus, each study would provide at most one effect size for each category. This approach would reduce the nonindependence of effect sizes, yet would not involve discarding data.

While this approach appears appropriate for use with main effects, its application to eliminating and interaction tests could be problematic. After collapsing within each category, it would no longer be possible to assign a value to each effect size along other dimensions, because the effect sizes had been collapsed across these other dimensions. Suppose, for example, that one desired to test the interaction between the source-of-outcome measure and the measurement technology of the outcome measure. It is likely that for some studies measurement technology would vary *within* the categories of source-of-outcome measure; e.g., one might have observers making behavioral ratings as well as behavioral counts, which represent two different levels of measurement technology. This would preclude tests involving measurement technology.

Consequently, our primary analyses were based on the intermediate approach we described above. However, we did feel that the issues related to nonindependence of effect sizes merited consideration. Consequently, we reran our validity analyses, applying these alternative strategies. To implement the data-dropping approach, we created two new datasets, each time randomly selecting from the dataset one effect size per study. Two different randomly selected datasets were used to determine the consistency of results produced by this random selection procedure. We also reran our main effect analyses collapsing up to and within the level of analysis. Because, as noted above, it was not feasible to test interaction or eliminating tests, we only tested main effects using this strategy. The results of these analyses, as well as those of the data-dropping approach, are reported at the end of our section on validity factors.

Classification and Coding Systems

Sample and treatment characteristic (e.g., type of treatment) coding systems were taken directly from our previous meta-analysis. Classification systems for the validity factors and type of control group are described below. Fifteen percent of the studies were randomly selected for independent coding by two judges; kappas (reported below) were all in the high end of the good to the excellent range (Fleiss, 1981).

Internal Validity Dimensions

Experimental Attrition. For each study, we computed the percentage of control group and the percentage of treatment group subjects who failed to complete the post-treatment assessment.

Subject Assignment. Studies were categorized as having assigned individual subjects to treatment and control groups: (1) nonrandomly; (2) randomly, without matching of treatment and control groups; (3) randomly, with matching of groups (via group means or pairs of individuals), on at least one dependent variable ($\kappa = 0.77$). We also assessed the impact of randomization and matching by comparing treatment and control groups' pretreatment status on the dependent measures. These analyses were performed on the subsample of studies which provided pre-treatment means and standard deviations for dependent variables.

Reactivity I: Measurement Technology. We coded along three dimensions relevant to reactivity or potential for subjective biasing in the outcome measures. The first of these dimensions was the outcome measurement technology. Like Shapiro and Shapiro (1982), we coded along a 1-to-4, "soft" to "hard" scale, with dependent measures lower on the scale deemed more susceptible to biasing. Self-reports of symptoms, attitudes or emotions, and projective tests were coded as 1; ratings based on subjective scales (e.g., *seldom, sometimes, frequently*) were coded as 2; count data (which included time, occurrence vs. nonoccurrence, number of occurrences, etc.) were coded as 3; life-event data such as arrests, GPA, etc., were coded as 4 ($\kappa = 0.89$).

Reactivity II: Rater Blindness. Our second reactivity dimension involved the experimental blindness of raters. For this factor, we coded whether outcome data came from an individual who was aware, or unaware, of the subject's experimental group assignment ($\kappa = 0.78$). In many instances, this information was directly noted in the research report. Self-reports were coded as missing in regards to rater blindness; while some researchers might consider the subject to be a rater, we felt that this form of assessment did not involve a "rater" in the traditional sense of the word. Data initially collected independent of the research project (e.g., attendance records, recidivism) was considered to have originated from a blind source. Although it is possible that awareness of group assignment could influence what was placed in a child's record, under such conditions (i.e., for data collected independently of the outcome investigation) it seemed unlikely or impossible that a rater would have known whether a child was assigned to a treatment or control group. .

Reactivity III: Subject Blindness. Rather than coding whether the subjects were aware of their own experimental group assignment, we coded whether they were aware that the outcome assessment was being made (κ

= 0.71), because (1) awareness that an outcome assessment was being made could well result in a child's acting in an unrepresentative or biased fashion (just as awareness of subject assignment might result in a rater making ratings in a biased fashion), and (2) subjects are seldom unaware of their experimental group status. If the outcome rating was based on 2 or more weeks of behavior, we considered the child to have been blind to the assessment because it seemed probable that if the rating was made over an extended period of time, it would have been difficult for the child to have consistently biased his or her behavior. Thus, for instance, most behavior checklists were considered subject-blind since the ratings were made on behavior over an extended period of time. For data collected independently of the research project (e.g., arrests, school suspensions), subjects were also considered to have been assessment-blind, since it again would have been difficult for the child to have consistently biased his or her behavior in response to the assessment.

External Validity Dimensions

As noted above, we categorized studies in regard to three dimensions relevant to the extent to which they approximated nonresearch clinical situations.

Treated Anyway. We first determined whether subjects would have been in some form of psychological or behavioral treatment irrespective of the research project (x = 1.00). We subsequently refer to this as the "treated anyway" factor.

Professional Therapist. Second, we distinguished between therapists whose principal vocational function was clinical (e.g., staff members of a clinic) from those whose function was not primarily clinical (e.g., faculty of psychology department, graduate students). We will refer to this as the "professional clinician" factor (x = 0.73).

Setting. Third, we categorized studies as to whether the therapy occurred in a clinical (e.g., hospital) or nonclinical (e.g., elementary school) setting (x = 0.82). We considered settings within academic psychology departments to be nonclinical since the main activity occurring in such settings was nonclinical.

We considered other potential external validity dimensions, such as flexibility of treatment administration and initial severity of disorder. However, based on conceptual considerations (e.g., substantial confounding of type of treatment with administration flexibility) and/or practical issues (e.g., difficulty in reliably coding initial severity), we chose to limit our analog dimensions to the three noted above. More importantly, many other poten-

tial dimensions seemed largely subsumed by these three dimensions (e.g., initial severity would likely be highly correlated with *treated anyway,* since children referred for clinical services should have more severe problems than those not referred).

Type of Control Group

Control groups were classified into one of four categories ($\kappa = 0.81$): (1) assessment only—subjects had no contact with the investigators except for pre- and post-assessments; (2) waiting list—subjects had no contact with investigators except for assessment, but were placed on a waiting list, or otherwise informed that they could receive services after the end of the control period; (3) attention-placebo—subjects had contact with the investigators, or with other persons at the instigation of the investigators, roughly equivalent to contact time with the treated subjects (this contact generally consisted of little more than social interaction, or undirected group interaction, and was designed to control for the extra attention that the treated children received); (4) minimal treatment—subjects received general, nonspecific (relative to the goals of the research project) treatment, in contrast to the experimental treatment; this often was a standard treatment administered by milieu staff at an in-patient setting.

We also investigated the potential impact of type of control group by assessing the different control groups' pre- to post-treatment improvement. These analyses were performed on the subsample of studies ($N = 26$) which provided pre- and post-treatment means and standard deviations on the dependent variables for the control group(s).

RESULTS

Overview of Analyses

In analyzing the eight validity factors, we used general linear model analyses (Hays, 1981). Attrition, measurement technology, and subject assignment were treated as continuous variables; "treated anyway," "professional therapist," setting, and rater and subject blindness were treated as categorical variables. We first tested each validity factor as a main effect, with outcome as the dependent variable. If a main effect was significant, we statistically eliminated (i.e., controlled for; see Appelbaum & Cramer, 1974) the other validity factors one at a time, to determine if this significant

main effect might be, in part, the result of confounding with another validity factor.

We next tested interactions; the purpose of these tests was to determine whether the validity factors influenced relations between other variables and outcome. Consequently, for these interaction tests we selected those substantive variables which in our previous meta-analysis (Weisz et al., 1987a) had shown a significant relation with outcome: age, type of treatment, and source-of-outcome measure (e.g., parents, teachers). Accordingly, we tested interactions between each of these variables and each validity factor.

Overall, this produced a moderately large number of significance tests, resulting in an increased probability of a Type I error. We initially considered applying the Bonferroni alpha level adjustment procedure to control for the number of significance tests that we were inspecting. However, the Bonferroni procedure becomes "intolerably conservative" (Hays, 1981) as the number of tests becomes moderate to large, increasing the probability of a Type II error. In the case of our validity factor analyses, we felt that a Type II error (i.e., deciding that a validity factor did not influence outcome when it in fact did) was at least as important to avoid as a Type I error (i.e., deciding that a validity factor did influence outcome when it in fact did not). Consequently, we did not use an alpha level correction procedure for these validity tests.

Internal Validity Dimensions

Attrition

Attrition rates for treatment and control groups differed marginally ($p < .10$), although the actual difference was less than 2 percentage points; on the average, 5.53% of treatment group subjects and 4.22% of control group subjects withdrew from their experiments prior to completion. Because the correlation between treatment and control group attrition was quite high ($r = .76, p < .0001$), we considered collapsing across treatment and control groups. However, it seemed plausible that the two variables might have different interactive relations (e.g., treatment attrition might interact with type of treatment, whereas control attrition might not). Consequently, we choose to test control group and treatment group attrition separately.

We tested the main effects of control and treatment group attrition on effect size; we then tested interactions of both kinds of attrition with the three substantive variables (age, type of treatment, source). All tests were nonsignificant, all $F < 2.75$, all $p > .10$. We also assessed the effect of

differential attrition (i.e., treatment minus control group attrition), again looking at its main and interactive effects. Once again, all tests were nonsignificant, all $F < 1.00$, all $p > .40$.

Subject Assignment

We performed two sets of analyses with our subject assignment variable. The first set involved tests of the main and interactive effects of subject assignment on post-treatment effect size; in the second we determined whether randomization and matching were effective in producing pretreatment treatment and control group equivalency vis-à-vis the dependent measures.

Subject Assignment Effects at Post-Treatment. We found a significant main effect of subject assignment on outcome ($F(1, 154) = 5.01, p < .03$), with higher levels of experimental rigor in subject assignment resulting in larger effect sizes (standardized beta = 0.18; see Table I for means). We next performed eliminating tests to determine if this significant main effect might in part have been the result of confounding with other validity factors. Since subject assignment was constant across all dependent measures within a study, in the preceding analysis effect sizes were collapsed across dependent measures within treatment group. However, collapsing was not

Table I. Mean Effect Sizes for Validity Factors

Factor	Level of factor	Effect size
Control group attrition		−0.06[a]
Treatment group attrition		−0.09[a]
Rater blindness	Yes	0.72
	No	0.68
Subject blindness	Yes	0.55
	No	0.73
"Treated anyway"	Yes	0.85
	No	0.77
"Professional clinician"	Yes	0.51
	No	0.87
Setting	Clinical	1.04
	Nonclinical	0.76
Subject Assignment	Nonrandom	0.34
	Random, no matching	0.80
	Random, Matching	1.04
Measurement technology	Level 1 (e.g., self-reports)	0.49
	Level 2 (e.g., subjective ratings)	0.45
	Level 3 (e.g., count data)	0.90
	Level 4 (e.g., arrests)	0.55

[a]The correlation between attrition and effect size is provided, rather than the mean size, because attrition was analyzed as a continuous variable.

appropriate for some of the validity factors to be eliminated (e.g., measurement technology) because they differed across dependent measures (e.g., within a particular study, different dependent measures sometimes used different measurement technologies). Thus, to be equivalent, all eliminating tests were run with uncollapsed effect sizes. This in turn suggested that we rerun our main effect test for subject assignment with uncollapsed effect sizes, so that we could have an equivalent test with which to compare to our eliminating tests.

When we used uncollapsed effect sizes, the subject assignment effect was significant ($F(1, 378) = 3.35$, $p < .001$). Although most of the other validity factors reduced the significance of subject assignment when eliminated, the only one to do so to a notable degree was setting, and even then subject assignment remained significant: $F(1, 331) = 4.45$, $p < .04$. Thus, the relation between outcome and randomization does not appear to have been the result of confounding with other validity factors.

Finally, we tested the interactions between subject assignment, and age, type of treatment, and outcome measure source. None of these tests was significant.

Subject Assignment Effects at Pre-Treatment. In the second set of subject assignment analyses, we assessed whether randomization and matching are, in fact, effective in producing groups that are similar on the dependent measures at pre-treatment. In these analyses, our dependent variable was an "effect size" based on pre-treatment means and pre-treatment control group standard deviation (i.e., mean of the treatment group minus the mean of the control group, divided by the control group standard deviation). These analyses were performed for the subsample of studies which provided pre-treatment means and standard deviations.

Treatment and control groups differed significantly at the $p < .05$ level at pre-treatment on 15.3% (27 of 177) of the dependent measures. On the average, control and treatment groups differed (irrespective of direction) by 0.44 of a standard deviation, which was significantly different from zero ($p < .0001$). When the direction of the difference was taken into account, the pre-treatment effect size was -0.10, indicating that prior to treatment, treatment groups averaged 0.10 of a standard deviation below (i.e., less well adjusted) the control group; this group difference was also significantly different from zero.

The fact that the average treated child began therapy one-tenth of a standard deviation below the average untreated child suggests that our original estimate of the effectiveness of therapy (in Weisz et al., 1987a) may actually have been a slight underestimate, since treated children had to improve one-tenth of a standard deviation more than control children just to be equal at post-treatment. To test whether this influenced our estimates of the relative efficacy of behavioral and nonbehavioral treatments, we computed the

pre-treatment comparisons separately for the two categories of therapy, and tested whether they differed from zero. Behavioral treatment groups were 0.11 ($p < .05$) of a standard deviation below their control groups at pre-treatment; nonbehavioral treatment groups were 0.04 (n.s.) of a standard deviation below their control groups. Thus, the efficacy of behavioral studies relative to nonbehavioral studies may actually be slightly greater than our original estimate (Weisz et al., 1987a), since behaviorally treated children were starting at a slight disadvantage, relative to nonbehaviorally treated children. Of course, this inference must be qualified by the fact that the comparison of behavioral and nonbehavioral therapies was not made directly in the same studies, but rather was based on the comparison of different groups of studies.

Does matching control and treatment groups pre-treatment at least partially resolve this problem of pre-treatment inequivalence? Apparently so. When treatment and control groups were not matched, they differed by 0.12 of a standard deviation ($p < .05$). When they were matched, they differed by 0.02 (n.s.) of a standard deviation.

Measurement Technology of Outcome Measures

The measurement technology of the outcome measure was significantly related to effect size ($F(1, 380) = 7.34$, $p < .01$), with "harder" outcome measures (e.g., counts, life-events) showing larger effects (standardized beta = 0.14; see Table I for means).[8] Since measurement technology was based on uncollapsed effect sizes, eliminating tests required no change vis-à-vis collapsing. Location, "treated anyway," subject assignment, attrition, and child blindness all reduced the significance of measurement technology slightly when eliminated. However, "professional clinician" reduced it to $p > .13$, and rater blindness to $p > .17$. Thus, there is some evidence suggesting that the significant effect of measurement technology was due in part to confounding with other validity factors.

Finally, we tested the interactions between measurement technology, and age, type of treatment, and source-of-outcome measure. All interactions were nonsignificant.

Rater Blindness

When we tested whether rater awareness of the subjects' experimental group status influenced effect size, we found a nonsignificant main effect ($p > .50$). All interaction tests were also nonsignificant.

[8]When analyzed as a categorical variable, measurement technology was significant ($F(3, 378) = 7.84$, $p < .0001$). Nonetheless, we chose to treat this variable as a continous factor, since it was conceptualized and coded as a continous variable.

Subject Blindness

The main effect of subject blindness on effect size was significant ($F(1, 351) = 4.07$, $p < .05$), with experimentally blind subjects producing smaller effect sizes (see Table I for means). With the exception of subject assignment, eliminating the other validity factors changed the significance of subject blindness only marginally. When subject assignment was eliminated, however, subject blindness was no longer significant ($F(1, 340) = 2.14$, $p < .15$). No interaction tests were significant.

External Validity Dimensions

"Treated Anyway"

We first tested the "treated anyway" factor (i.e., whether subjects would have been in treatment irrespective of the research project). The main effect for this factor was nonsignificant ($F(1, 161) = 0.21$, $p > .50$; see Table I for means), as were the interactions with age, type of treatment, and source-of-outcome measure.

"Professional Clinician"

We next tested the main effect for "professional clinician" (i.e., whether the principal vocational function of the individual administering the treatment was clinical). This variable's main effect was nonsignificant ($F(1, 137) = 2.02$, $p > .15$; see Table I for means), as were all interactions.

Setting

Our last external validity factor was setting (i.e,, whether treatment took place in a clinic). The main effect of this variable was also nonsignificant ($F(1, 143) = 1.74$, $p > .18$; see Table I for means), as were all interactions.

Alternative Analytic Strategies

As noted above, most studies in our pool initially produced a number of effect sizes, resulting in two potential problems: (1) disproportionate weighting of those studies with the most measures, possibly resulting in a biased estimate of the effect size; and (2) nonindependence of observations, possibly resulting in an underestimation of the error variance (Strube & Hartmann, 1983). Although we felt that our collapsing strategy was most appropri-

ate (see above), we also wanted to empirically determine the consequences of using multiple-outcome measures, and evaluate the random selection of effect sizes strategy, as well as the collapsing up to and within the level of analysis approach.

We first assessed the relation between effect size and the number of effect sizes within a study, to determine if disproportionate weighting may have biased our estimate of the effect size. There was a significant relation, $r = -.29, p < .0001$ between number of effect sizes and within-study effect size, with studies containing more effect sizes producing *smaller* effect sizes. Thus, if anything, disproportionate weighting led to an underestimate of the effect size.

We next determined whether the effect size variance may have been attenuated by nonindependence of the effect sizes. When uncollapsed across all dimensions, the variance for the effect size equaled 0.70; when collapsed across all dimensions within-study, the (between study) variance equaled 1.02. Within-study effect size variability equaled 0.45, which differed significantly from the between-study variance ($F(104, 90) = 2.27, p < .0001$).

Finally, we assessed our two supplementary analytic strategies. To test the random selection of one effect size per study strategy, we created two datasets, each time randomly selecting one effect size from each study. As Table II indicates, of the three main effects significant in the primary analyses, in each of the two random datasets one effect was significant and one effect was marginally significant; which effect was significant or marginal differed across the two random datasets, however. As Table II also indicates, of the three main effects significant in the primary analyses, two were significant using the collapsing up to and within the level of analysis strategy. Whereas no interactions (of 27) were significant in the primary analyses, two

Table II. Significance of Main Effects for Validity Factors, Under Different Analytic Approaches

	Analytic approach[a]		
Validity factor	Primary	Data drop	Collapse within
Attrition, control group	n.s.	n.s./n.s.	n.s.
Attrition, treatment group	n.s.	n.s./n.s.	n.s.
Measurement technology	s.	m.g./s.	s.
"Professional clinician"	n.s	n.s./n.s.	n.s.
Rater blindness	n.s.	n.s./n.s.	n.s.
Setting	n.s.	n.s./n.s.	n.s.
Subject assignment	s.	n.s./m.g.	n.s.
Subject blindness	s.	s./n.s.	s.
"Treated anyway"	n.s.	n.s./n.s.	n.s.

[a]*Note: Primary* refers to the primary analyses; *data drop* refers to the analyses based on the two randomly selected datasets; *collapse within* refers to the strategy based on collapsing up to and within the level of analysis. s = significant; m.g. = marginally significant; n.s. = nonsignificant.

interactions were significant in one randomly selected dataset but not in the other. As noted previously, interaction tests were not feasible using the collapsing up to and within the level of analysis approach.

In sum, these results suggest that the significance of the main effect for subject assignment, and to a lesser extent child blindness and measurement technology, should be approached with a measure of caution. These results also suggest that findings based on one randomly selected effect size per study may depend in part on the particular random selection that is made, and that randomly selected datasets may be as different from one another in the effects they generate as they are from the complete sample.

Overall Effect of Validity Factors

We determined the overall effect of the validity factors by computing a multiple-regression equation with effect size as the criterion, and "treated anyway," "professional therapist," setting, attrition, subject assignment, measurement technology, and rater and subject blindness as the predictors. These variables produced an adjusted (for degrees of freedom) R^2 of .07, with $F(8, 188) = 2.73$, $p < .008$, suggesting that as whole, the validity factors account for a small but statistically significant amount of outcome variance. By way of comparison, a model containing the seven substantive main effects from our previous meta-analysis (age, type of treatment, type of problem, professional status, group vs. individual treatment, content and source-of-outcome measures) accounted for 11% of the variance in outcome after adjustment for degrees of freedom.

Subject assignment and subject blindness made significant ($p < .05$) unique contributions to the validity predictors model, while measurement technology made a marginally significant unique contribution ($p < .10$). These three factors were also significantly correlated with outcome: for measurement technology, $r = .14$; for subject assignment, $r = .17$; for subject blindness, $r = -.11$, with positive correlations indicating that increasing experimental rigor was associated with an increasing effect size. All other correlations between validity factors and outcome were nonsignificant.

This multiple-regression equation also allowed us to determine the magnitude of the effect size one might expect if one were able to conduct what might be called the "methodologically ideal" experiment: one where subjects were clinic-referred, treated by a professional in a clinic, with randomization, matching and no attrition, blind raters and subjects, and assessment based on the least reactive outcome measures (i.e., life-event data). This estimate was obtained by using the model parameter estimates derived from the regression equation, and then computing the predicted value for a study having the "most valid" level of each validity factor. The predicted effect size for this ideal model was 1.14. This suggests that *if* one were able to run

such a study as described above, it would produce an effect size substantially larger than that produced by the average study in our sample (0.79). Thus, if anything, it would appear that methodological inadequacies have lowered estimates of the effectiveness of therapy.

Type of Control Group

Control Group Effects at Post-Treatment

We next sought to determine: (1) whether effect size was constant across type of control group in general; and (2) more specifically, whether attention-placebo groups produced different effect sizes than other control groups. Outcome (i.e., post-treatment effect size collapsed across dependent measures) served as the dependent variable and type of control group as the independent, categorical variable. The ANOVA for this test was based on three contrasts chosen to represent questions pertaining to attention-placebo groups and nonspecific therapy effects; each contrast represented one degree of freedom in the overall 3-degree-of-freedom test for type of control group. These contrasts were (1) attention-placebo minus the mean of assessment only and waiting lists; (2) attention-placebo minus assessment only; and (3) minimal treatment minus the mean of the other three groups. Although the choice of contrasts has no effect on the overall F so long as none of the contrasts are linearly dependent (Hays, 1981), inspection of the individual contrasts allowed us to determine whether the attention-placebo strategy does in fact control for nonspecific factors. (This is based on the supposition that if the attention-placebo strategy does control for these factors, studies utilizing such control group should produce smaller effect sizes than studies that do not, since these nonspecific effects are, in essence, being removed from the effect size.)

The overall test for type of control group was nonsignificant ($F < 1.25$, $p > .30$), as were the tests for the three contrasts (all $p > .15$). Thus, these analyses all suggest that the type of control group experimenters chose did not influence the magnitude of the effect that they find. Nor do these findings support the efficacy of attention-placebo groups. In fact, attention-placebo studies actually produced a nonsignificantly *larger* effect than those using untreated controls (1.09 vs. 0.79, collapsing across assessment-only and waiting list groups).

Control Group Improvement

Our second set of control group analyses was designed to assess within control group improvement, pre- to post-treatment. We hoped to determine whether different control groups improved (or perhaps deteriorated) at differ-

ent rates; if attention-placebo groups do involve certain nonspecific effects, one would expect that they should show higher rates of improvement (or lower rates of deterioration) than assessment-only or waiting list controls.

In this set of analyses, we analyzed those control groups within our sample for which we had sufficient information (i.e., pre- and post-treatment means and standard deviations) to make *within* control group, time 1 vs. time 2, comparisons. For these tests, we computed an effect size based on each control group's pre-treatment score minus its post-treatment score (or vice versa, depending on the direction of the outcome measure), divided by its pre-treatment standard deviation.[9] Thus, we assessed each control group's improvement relative to itself.

Within each study, we collapsed across different dependent measures. This produced a sample size of 26 comparisons (i.e., effect sizes) based on 494 subjects; see Table III for frequencies by type of control groups. The mean effect size across the various controls groups was 0.31, which differed significantly from zero ($t(25) = 2.68$, $p < .02$). Thus, in this subsample, control groups showed significant improvement from pre- to post-treatment. The control groups did not differ significantly in regards to improvement ($F(3, 19) = 0.96$, $p > .40$; see Table III), even when a low-frequency cell (minimal treatment) was dropped. These tests were of relatively low power, however. It should be noted that when effect sizes were not collapsed across dependent variables, the groups did differ significantly, with waiting list controls showing significantly more improvement than the other groups. (The difference between the results of these two tests was due primarily to the increase in degrees in freedom, rather than a change in the cell means.)

Changes in Variability

To determine whether therapy results in increased behavioral variability, we focused on those studies which reported the variances (or standard deviations) of the outcome measures at both pre- and post-treatment. For each outcome measure, we first computed a *t* statistic for the heterogeneity of nonindependent variances (Howell, 1982), separately for the treatment and control groups, pre- vs. post-treatment; we then converted these *t*s to effect sizes. Uncollapsed across dependent measures within a study, both treat-

[9]We chose this definition ($[C_1 - C_2]$) / $SD[C_1]$) for a change effect size, rather than Cohen's (1977) change effect size ($d_t' = [C_1 - C_2]/SD[C_1 - C_2]$), because we felt the former was more comparable to the standard Glass effect size ($d_G = [X_1 - X_2])/SD[X_1]$) as well as to Cohen's standard effect size ($d = [X_1 - X_2])$/pooled $SD[X_1, X_2]$). Cohen's d_t' is based on the *sum* of variances ($SD[C_1 - C_2] = [\mathrm{VAR}[C_1] + \mathrm{VAR}[C_2] - 2\mathrm{COV}[C_1, C_2]]^{1/2}$), whereas d_G is based on a single group's variance ($SD[X_1] = \mathrm{VAR}[X_1]^{1/2}$) and d is based on the *mean* of two variances (pooled $SD = \{[\mathrm{VAR}(X_1) + \mathrm{VAR}(X_2)]/2\}^{1/2}$, for equal-sized groups). We did, however, analyze control group change using d_t', and found that $d_t' = 0.20$ ($t(25) = 2.42$, $p < .05$)). As in the other analysis, the control groups did not differ in regard to the amount of change.

Table III. Pre- to Post-Treatment Effect Sizes, by Type of Control Group

Type of control group[a]	*N*	Mean *ES*
Assessment-only	8	0.22
Waiting list	5	0.48
Attention-placebo	8	0.06
Minimal treatment	2	0.07
Average		0.31

[a]*Note:* Three studies were uncodable for control group.

ment and control group variability increased significantly ($t(180) = 4.53$, $p < .0001$, mean $ES = 0.91$; $t(180) = 4.11$, $p < .0001$, mean $ES = 0.42$, respectively); the treatment group increase was significantly greater than that of the control group ($t(180) = 2.20$, $p < .05$). When collapsed across dependent measures, only the treatment group showed increased variability ($t(25) = 2.29$, $p < .05$, mean $ES = 1.24$).

We next tested whether treatment and control group variances differed significantly post-treatment; uncollapsed, they did ($t(1.80) = 2.62$, $p < .01$, $ES = 0.22$). However, although the mean effect size increased, when collapsed across outcome measures the difference was no longer significant ($t(25) = 1.31$, $p < .25$, $ES = 0.30$), even when a directional, one-tailed test was used.

DISCUSSION

How much impact do validity factors have on the results of child psychotherapy research? At least in our data, almost two-thirds as much as the substantive factors assessed in our previous meta-analysis: Whereas these substantive factors accounted for 11% of outcome variance (after adjustment for degrees of freedom), our validity factors accounted for 7%. This suggests that relative to clinical or child-characteristic factors, the kinds of variables analyzed in the present investigation may have a substantial (though somewhat smaller—i.e., two-thirds as strong) impact on outcome.

None of our validity interaction tests were significant, however, and we believe that such tests may be more important than tests of main effects. Validity factors functioning as main effects could lead to over- or underestimates of therapeutic effectiveness; interaction effects could lead to incorrect conceptualizations of relations between variables. Despite the large number of interactions we tested, we found little evidence that variations in validity factors across studies have created distorted pictures of relations.

This is of course not justification for ignoring these factors, since there are other variables which we did not test with which these factors might interact.

Subject assignment—i.e., randomization and matching—appears to be the validity factor exerting the strongest influence on effect size. The magnitude of its impact was in part due to the fact that, while overall, control and treatment groups showed substantial pre-treatment differences (i.e., groups on the average differed by almost half a standard deviation, with over 15% of these differences significant), groups that were matched showed smaller, nonsignificant treatment/control differences. This suggests that some form of pre-treatment matching may be advisable.

Measurement technology of the outcome measure was the second most influential validity factor. As was the case with subject assignment, increased experimental rigor was related to larger effect sizes. For the third most influential validity factor, subject blindness, increased experimental rigor was related to smaller effect sizes. However, taken together these results, in conjunction with our findings regarding the overall effect of the validity factors, argue against the hypothesis that methodologically weak studies have led to an overestimation of the effectiveness of therapy. Indeed, the findings suggest that the more rigorous a pool of studies one selects, or the more rigorous a study one designs, the more substantial an effect size one is likely to find.

Although subjects who were not experimentally blind produced larger effect sizes than those who were, observers (e.g., teachers, researcher observers) appeared uninfluenced by experimental blindness. This suggests that subject blindness may be more important to mantain than observer blindness, which may be useful to remember during the design planning stages of outcome studies.

Our external validity factors appeared to influence outcome relatively little; of the 12 tests (i.e., three main effects and nine interactions), none were significant. This may be encouraging news to researchers attempting to redress the problem of design limitation tradeoffs, since such tradeoffs (particularly between internal and external validity; Kazdin, 1978b) may be an unavoidable part of clinical research. For example, if one has an analog sample, generalizability may be unclear. To deal with this problem, we have suggested (Weisz & Weiss, 1989b) the use of a form of design triangulation. By combining the results of individual studies, each with a slightly different empirical perspective on child psychotherapy, each with its view partially blocked by its design limitations, one may begin to piece together a picture of the whole. The lack of influence of the external validity factors holds out the hope that these different research approaches are at least focusing on the same basic process.

The fact that the external validity factors appeared to have little influence on outcome might suggest to some that researchers should restrict their efforts to analog samples, where it is relatively easy to control internal

validity. We believe this would be a mistake, however: Remember that we tested the interaction between our analog factors and *three* variables. There are obviously many other variables with which the analog factors might interact. We failed to reject the null hypothesis for only these variables.

All of the preceding results were based on our intermediate strategy in regard to the unit of analysis—i.e., collapsing across dependent measures when appropriate. Results of our alternative strategies—random selection of single effect size per study, and collapsing up to and within the level of analysis—produced results which were partially inconsistent, both in comparison to results based on the original, complete dataset, as well as across alternative strategies. Although this might seem to suggest that the complete dataset it inconsistent, it is important to recall that the original dataset was consistent enough to produce several significant main effects. Given the disagreement between randomly selected datasets, and the unfeasibility of computing interactions with the collapsing within the level of analysis strategy, we felt it advisable, at least in this dataset, to focus on the primary analyses. However, the results of the alternative analytic strategies do suggest that the significance of the main effects for subject assignment, and to a lesser extent child blindness and measurement technology, should be interpreted with a measure of caution.

In the complete dataset, we found no evidence suggesting that our previous positive conclusions (Weisz et al., 1987a) regarding the efficacy of child psychotherapy must be qualified with respect to the type of control group employed: Outcome appears to be relatively unaffected by the type of control group chosen by the experimenter. Thus, it appears valid to compare and aggregate studies using different control groups.

Further, our findings support Kazdin's (1978b) concern that the attention-placebo control strategy may not achieve its stated objectives. Treatment groups compared to attention-placebo control groups actually showed nonsignificantly *larger* effect sizes than those compared to no contact controls. If the attention-placebo groups were in fact controlling for nonspecific factors, one would have expected that treatment groups compared to them would have shown smaller effect sizes. And in a separate analysis, attention-placebo groups showed nonsignificantly *less* improvement from pre- to post-treatment than no-contact controls.

Our results regarding control group improvement may also have important implications for another type of control group, the "dropout" or "therapy refusers" comparison group. The use of such a control group, composed of children who complete clinic intake procedures but never return for therapy, has been proposed by some researchers (e.g., McAdoo & Roeske, 1973; Weisz & Weiss, 1989a; Weisz, Weiss, & Longmeyer, 1987b) as one possible strategy for obtaining a control group in clinics where random assignment to a no-treatment control group would be difficult or impossible.

One potential problem with such a strategy is that the "dropouts" and "continuers" (i.e., the treatment group) might differ on such important variables as prognosis. In fact, in a recent outcome investigation based on this strategy (Weisz & Weiss, 1989a), we found that a dropout control group did show significant improvement; one possible explanation for this finding is that the children who dropped out did so because they had a better prognosis than the children who remained. If true, this would invalidate the dropouts as a control group. However, our present finding that randomized control groups appear to improve across time suggests that the improvement noted in the dropout control group of Weisz and Weiss (1989a) may not be unique to dropout controls. In fact, the magnitude of improvement was quite comparable across studies: 0.31 in the present study, 0.39 in Weisz and Weiss (1989a).

Among the treated groups, we found that outcome measure variance increased significantly from pre- to post-treatment. This suggests that when evaluating the results of child therapy, it may be worthwhile to consider changes in variability as well as changes in the mean. Treatment and control group post-treatment variances differed significantly, though only when outcome measures were not collapsed; this is suggestive, but does not argue unequivocally against pooling. Still, the findings may be sufficiently strong to make pooling inadvisable, particularly since inappropriate pooling could lead to biased estimates of the parameter (i.e., the effect size), whereas failure to pool when appropriate could lead to inefficient parameter estimates; we see bias as more problematic than inefficiency. At a minimun, if meta-analysts intend to pool, they should first compare the variability of their groups.

There are several limitations of these analyses which it is important to note. First, our tests of interactions with the validity factors were only two-way interactions; decreasing cell sizes precluded higher-level interactions. Thus, it is possible that the validity factors would have had more of an impact if these more complex relations could have been assessed. Hopefully, as the number of studies in this area increases, it will become possible to probe for more complex relations of this type.

As in any meta-analysis, there was substantial confounding of independent variables. Meta-analysis is inevitably a correlational technique; the variables under analysis are not controlled by the individual performing the analyses, and different conditions, across studies, are not randomly assigned. Although we attempted to control for this through our eliminating tests, such *post hoc* adjustments can never be totally successful at unconfounding variables.

As we noted above, Hunter, Schmidt, and Jackson (1982) have described a series of techniques designed to adjust the effect size and its variance for various artifacts. Use of these procedures would allow researchers to address a variety of interesting questions, such as whether effect size variabili-

ty can be explained by study artifacts. Unfortunately, we were unable to apply these procedures because the large majority of outcome study reports in our sample did not include the information necessary to compute these adjustments.

Consequently, we feel that it is important to urge authors, and editors, that certain data always be included in primary research reports. At a minimum, reports should include post-treatment means and standard deviations for treatment and control groups for all outcome measures, *regardless of whether there is a significant effect for the measure.* Likewise, reliability estimates for the measures should be provided. Pre-treatment means and standard deviations are also important to include, in that they allow for assessment of the impact of therapy on variability, and for evaluation of changes in untreated groups. Including such information should allow a study to provide maximal contribution to the literature.

APPENDIX

Studies included in the meta-analysis, post-treatment as well as followup assessments.

Alexander, J. F., & Parsons, B. V. (1973). Short term behavioral intervention with delinquent families: Impact on family process and recidivism. *Journal of Abnormal Psychology, 81,* 219-225.

Alper, T. G., & Kranzler, G. D. (1968). A comparison of the effectiveness of behavioral and client-centered approaches for the behavior problems of elementary school children. *Elementary School Guidance and Counseling, 3,* 35-43.

Andrews, W. R. (1971). Behavioral and client-centered counseling of high-school underachievers. *Journal of Counseling Psychology, 18,* 93-96.

Bandura, A., Grusec, J. E., & Menlove, F. L. (1967). Vicarious extinction of avoidance behavior. *Journal of Personality and Social Psychology, 5,* 16-23.

Bandura, A., & Menlove, F. L. (1968). Factors determining vicarious extinction of avoidance behavior through symbolic modeling. *Journal of Personality and Social Psychology, 8,* 99-108.

Baymur, F. B., & Patterson, C. H. (1960). A comparison of three methods of assisting underachieving high school students. *Journal of Counseling Psychology, 7,* 83-90.

Bean, A. W., & Roberts, M. W. (1981). The effect of time-out release contingencies on changes in child noncompliance. *Journal of Abnormal Child Psychology, 9,* 95-105.

Bender, N. N. (1976). Self-verbalization versus tutor-verbalization in modifying impulsivity. *Journal of Educational Psychology, 68,* 347-354.

Block, J. (1978). Effects of a rational-emotive mental health program on poorly achieving, disruptive high school students. *Journal of Counseling Psychology, 25,* 61-65.

Camp, B. W., Blom, G. E., Herbert, F., & Van Doornick, W. J. (1977). "Think aloud": A program for developing self-control in young aggressive boys. *Journal of Abnormal Child Psychology, 5,* 157-169.

Clement, P. W., & Milne, D. C. (1967). Group play therapy and tangible reinforcers used to modify the behavior of eight-year-old boys. *Behavior Research and Therapy, 5,* 301-312.

Cohen, N. J., Sullivan, J., Minde, K., Novak, C., & Helwig, C. (1981). Evaluation of the relative effectiveness of methylphenidate and cognitive behavior modification in the treatment of kindergarten-aged hyperactive children. *Journal of Abnormal Child Psychology, 9,* 43-54.

Cradock, C., Cotler, S., & Jason, L. A. (1978). Primary prevention: Immunization of children for speech anxiety. *Cognitive Therapy and Research, 2,* 389-396.

Cullinan, D., Epstein, M. H., & Silver, L. (1977). Modification of impulsive tempo in learning disabled pupils. *Journal of Clinical Psychology, 5,* 437-444.

Deffenbacher, J. L., & Kemper, C. C. (1974). Counseling test-anxious sixth-graders. *Elementary School Guidance and Counseling, 9,* 22-29.

Diament, C., & Colletti, G. (1978). Evaluation of behavioral group counseling for parents of learning disabled children. *Journal of Abnormal Child Psychology, 6,* 385-400.

Dorfman, E. (1958). Personality outcomes of client-centered child py.
therapy. *Psychological Monographs: General and Applied, 72* (XIII, No.
Whole No. 456).

Douglas, V. I., Parry, P., Marton, P., & Garson, C. (1976). Assessment of a cognitive training program for hyperactive children. *Journal of Abnormal Child Psychology, 4,* 389-410.

Durlak, J. A. (1977). Description and evaluation of a behaviorally oriented school-based preventive mental health program. *Journal of Consulting and Clinical Psychology, 45,* 27-33.

Egeland, B. (1974). Training impulsive children in the use of more efficient scanning techniques. *Child Development, 45,* 165-171.

Elliott, C. D., & Pumfrey, P. D. (1972). The effects of non-directive play therapy on some maladjusted boys. *Education Research, 14,* 157-161.

Evers-Pasquale, W., & Sherman, M. (1975). The reward value of peers: A variable influencing the efficacy of filmed modeling in modifying social isolation in preschoolers. *Journal of Abnormal Child Psychology, 3,* 179-189.

Feindler, E. L. (1984). Group anger control training for junior high school delinquents. *Cognitive Therapy and Research, 8,* 299-311.

Finch, A. J., Jr., Wilkinson, M. D., Nelson, W. M., III, & Montgomery, L. E. (1975). Modification of an impulsive cognitive tempo in emotionally disturbed boys. *Journal of Abnormal Child Psychology, 3,* 49-52.

Finney, B. C., & van Dalsem, E. (1969). Group counseling for gifted underachieving high school students. *Journal of Counseling Psychology, 16,* 87-94.

Fo, W. S. O., & O'Donnell, C. R. (1974). The buddy system: Relationship and contingency conditions in a community intervention program for youth with nonprofessionals as behavior change agents. *Journal of Consulting and Clinical Psychology, 42,* 163-169.

Forman, S. (1988). A comparison of cognitive training and response cost procedures in modifying aggressive behavior of elementary school children. *Behavior Therapy, 11,* 594-688.

Furman, W., Rahe, D. F., & Hartup, W. W. (1979). Rehabilitation of socially withdrawn preschool children through mixed-age and same-age socialization. *Child Development, 50,* 915-922.

Glenwick, D. S., & Barocas, R. (1979). Training impulsive children in verbal self-control by use of natural change agents. *Journal of Special Education, 13,* 387-398.

Graziano, A. M., & Mooney, K. C. (1980). Family self-control instruction for children's nighttime fear reduction. *Journal of Consulting and Clinical Psychology, 48,* 206-213.

Gresham, F. M., & Nagle, R. J. (1980). Social skills training with children: Responsiveness to modeling and coaching as a function of peer orientation. *Journal of Consulting and Clinical Psychology, 48,* 718-729.

Guerney, B. G., Jr., & Flumen, A. B. (1970). Teachers as psychotherapeutic agents for withdrawn children. *Journal of School Psychology, 8,* 107-113.

Gundel, R. C. (1981). The interaction of locus of control with three behavioral procedures in the modification of disruptive behavior in emotionally disturbed boys. *Multivariate Experimental Clinical Research, 5,* 99-188.

Hansen, J. C., Niland, T. M., & Zani, L. P. (1969). Model reinforcement in group counseling with elementary school children. *Personnel and Guidance Journal, 46,* 741-744.

Harris, K. R., & Brown, R. D. (1982). Cognitive behavior modification and informed teacher treatments for shy children. *Journal of Experimental Education, 58,* 137-143.

Harris, M. B., & Trujilla, A. E. (1975). Improving study habits of junior high school students through self-management versus group discussion. *Journal of Counseling Psychology, 22,* 513-517.

Heaton, R. C., Safer, D. J., Allen, R. P., Spinnato, N. C., Sr., & Prumo, F. M. (1976). A motivational environment for behaviorally deviant junior

high school students. *Journal of Abnormal Child Psychology, 4,* 263-275.

Hendrix, C. E., & Heckel, R. V. (1982). The effects of a behavioral approach on modifying social behavior in incarcerated male delinquents. *Journal of Clinical Psychology, 82,* 77-83.

Hinds, W. C., & Roehlke, H. J. (1970). A learning theory approach to group counseling with elementary school children. *Journal of Counseling Psychology, 17,* 49-55.

Jakibchuk, Z., & Smeriglio, V. L. (1976). The influence of symbolic modeling on the social behavior of preschool children with low levels of social responsiveness. *Child Development, 47,* 838-841.

Johnson, T., Tyler, V., Jr., Thompson, R., & Jones, E. (1971). Systematic desensitization and assertive training in the treatment of speech anxiety in middle-school students. *Psychology in the Schools, 8,* 263-267.

Kanfer, F. H., Karoly, P., & Newman, A. (1975). Reduction of children's fear of the dark by competence-related and situtational threat-related verbal cues. *Journal of Consulting and Clinical Psychology, 43,* 251-258.

Keller, M. F., & Carlson, P. M. (1974). The use of symbolic modeling to promote social skills in preschool children with low levels of social responsiveness. *Child Development, 45,* 912-919.

Kendall, P. C. (1982). Individual versus group cognitive-behavioral self-control training: One-year follow-up. *Behavior Therapy, 13,* 241-247.

Kendall, P. C., & Zupan, B. A. (1981). Individual versus group application of cognitive-behavioral self-control procedures with children. *Behavior Therapy, 12,* 344-359.

Kent, R. N., & O'Leary, K. D. (1976). A controlled evaluation of behavior modification with conduct problem children. *Journal of Consulting and Clinical Psychology, 44,* 586-596.

Kettlewell, P. W., & Kausch, D. F. (1983). The generalization of the effects of a cognitive-behavioral treatment program for aggressive children. *Journal of Abnormal Child Psychology, 11,* 101-114.

Kornhaber, R. C., & Schroeder, H. E. (1975). Importance of model similarity on extinction of avoidance behavior in children. *Journal of Consulting and Clinical Psychology, 43,* 681-607.

Kranzler, G. D., Mayer, G. R., Dyer, C. O., & Munger, P. F. (1966). Counseling with elementary school children: An experimental study. *Personnel and Guidance Journal, 44,* 944-949.

Krop, H., Calhoon, B., & Verrier, R. (1971). Modification of the "self-concept" of emotionally disturbed children by covert reinforcement. *Behavior Therapy, 2,* 201-294.

Ladd, G. W. (1981). Effectiveness of a social learning method for enhancing children's social interaction and peer acceptance. *Child Development, 52,* 171-178.

LaGreca, A. M., & Santogrossi, D. A. (1980). Social skills training with elementary school students: A behavioral group approach. *Journal of Consulting and Clinical Psychology, 48,* 220-227.

Laxer, R. M., Quarter, J., Kooman, A., & Walker, K. (1969). Systematic desensitization and relaxation of high-test-anxious secondary school students. *Journal of Counseling Psychology, 16,* 446-451.

Laxer, R. M., & Walker, K. (1970). Counter-conditioning versus relaxation in desensitization of test anxiety. *Journal of Counseling Psychology, 17,* 431-436.

Leal, L. L., Baxter, E. G., Martin, J., & Marx, R. W. (1981). Cognitive modification and systematic desensitization with test anxious high school students. *Journal of Counseling Psychology, 28,* 525-528.

Leitenberg, H., & Callahan, E. J. (1973). Reinforced practice and reduction of different kinds of fears in adults and children. *Behavior Research and Therapy, 11,* 19-30.

Lewis, S. (1974). A comparison of behavior therapy techniques in the reduction of fearful avoidance behavior. *Behavior and Therapy, 5,* 648-655.

Mann, J., & Rosenthal, T. L. (1969). Vicarious and direct counterconditioning of test anxiety through individual and group desensitization. *Behavior Research and Therapy, 7,* 359-367.

Massimo, J. L., & Shore, M. F. (1963). The effectiveness of a comprehensive, vocationally oriented psychotherapeutic program for adolescent deliquent boys. *American Journal of Orthopsychiatry, 33,* 634-642.

Mayer, G. R., Kranzler, G. D., & Matthes, W. A. (1966). Elementary school counseling and peer relations. *Personnel and Guidance Journal, 44,* 360-365.

McBrien, R. J., & Nelson, R. J. (1972). Experimental group strategies with primary grade children. *Elementary School Guidance and Counseling, 6,* 178-184.

McCollum, P. S., & Anderson, R. P. (1974). Group counseling with reading disabled children. *Journal of Counseling Psychology, 21,* 158-165.

Meichenbaum, D. H., & Goodman, J. (1971). Training impulsive children to talk to themselves: A means of developing self-control. *Journal of Abnormal Psychology, 77,* 115-126.

Miller, L. C., Barrett, C. L., Hampe, E., & Noble, H. (1972). Comparison of reciprocal inhibition, psychotherapy and waiting list control for phobic children. *Journal of Abnormal Psychology, 79,* 269-279.

Milos, M. E., & Reiss, S. (1982). Effects of three play conditions on separation anxiety in young children. *Journal of Consulting and Clinical Psychology, 50,* 389-395.

Miran, M., Lehrer, P. M., Koehler, R., & Miran, E. (1974). What happens when deviant behavior begins to change? The relevance of a social systems approach for behavioral programs with adolescent. *Journal of Community Psychology, 2,* 370-375.

Moulin, E. K. (1970). The effects of client-centered group counseling using play media on the intelligence, achievement and psycholinguistic abilities of underachieving primary school children. *Elementary School Guidance and Counseling, 5,* 85-98.

Muller, S. D., & Madsen, C. H., Jr. (1970). Group desensitization for "anxious" children with reading problems. *Psychology in the Schools, 7,* 184-189.

Murphy, C. M., & Bootzin, R. R. (1973). Active and passive participation in the contact desensitization of snake fear in children. *Behavior Therapy, 4,* 203-211.

O'Connor, R. D. (1969). Modification of social withdrawal through symbolic modeling. *Journal of Applied Behavior Analysis, 2,* 15-22.

O'Connor, R. D. (1972). Relative efficacy of modeling, shaping and the combined procedures for modification of social withdrawal. *Journal of Abnormal Psychology, 79,* 327-334.

O'Leary, K. D., Pelham, W. E., Rosenbaum, A., & Price, G. H. (1976). Behavioral treatment of hyperkinetic children. *Clinical Pediatrics, 15,* 518-524.

Ollendick, T. H., & Hersen, M. (1979). Social skills training for juvenile delinquents. *Behaviour Research and Therapy, 17,* 547-554.

Ostrom, T. M., Steele, C. M., Rosenblood, L. K., & Mirels, H. L. (1971). Modification of delinquent behavior. *Journal of Applied Social Psychology, 1,* 118-136.

Palkes, H., Stewart, M., & Freeman, J. (1972). Improvement in maze performance of hyperactive boys as a function of verbal training procedures. *Journal of Special Education, 5,* 337-342.

Palkes, H., Sterwart, M., & Kahana, B. (1968). Porteus maze performance of hyperactive boys after training in self-directed verbal commands. *Child Development, 39,* 817-826.

Peed, S., Roberts, M., & Forehand, R. (1977). Evaluation of the effectiveness of a standardized parent training program in altering the interaction of mothers and their noncompliant children. *Behavior Modification, 1,* 323-350.

Persons, R. W. (1966). Psychological and behavioral change in delinquents following psychotherapy. *Journal of Clinical Psychology, 22,* 337-340.

Pitkanen, L. (1974). The effect of simulation exercises on the control of aggressive behavior in children. *Scandinavian Journal of Psychology, 15,* 169-177.

Quay, H. C., Glavin, J. P., Annesley, F. R., & Werry, J. S. (1972). The modification of problem behavior and academic achievement in a resource room. *Journal of School Psychology, 10,* 187-198.

Randolph, D. L., & Hardage, N. C. (1973). Behavioral consultation and group counseling with potential dropouts. *Elementary School Guidance and Counseling, 7,* 284-289.

Randolph, D. L., & Wallin, K. R. (1973). A comparison of behavioral consultation with model reinforcement group counseling for children who are consistently off-task. *Journal of Education Research, 67,* 183-187.

Rickel, A. U., & Lampi, L. (1981). A two-year follow-up study of a preventive mental health program for preschoolers. *Journal of Abnormal Child Psychology, 9,* 455-464.

Ridberg, E. H., Parke, R. D., & Hetherington, E. M. (1971). Modification of impulsive and relective cognitive styles through observation of film-mediated models. *Developmental Psychology, 5,* 369-377.

Ritter, B. (1968). The group desensitization of children's snake phobias using vicarious and contact desensitization procedures. *Behavior Research and Therapy, 6,* 1-6.

Roberts, M. W., Hatzenbeuhler, L. C., & Bean, A. W. (1981). The effects of differential attention and time out on child noncompliance. *Behavior Therapy, 12,* 93-99.

Roberts, M. W., McMahon, R. J., Forehand, R., & Humphreys, L. (1978). The effect of parental instruction giving on child compliance. *Behavior Therapy, 9,* 793-798.

Sarason, I. G., & Ganzer, V. J. (1973). Modeling and group discussion in the rehabilitation of juvenile delinquents. *Journal of Counseling Psychology, 28,* 442-449.

Schlichter, K. J., & Horan, J. J. (1981). Effects of stress inoculation on the anger and aggression management skills of institutionalized juvenile delinquents. *Cognitive Therapy and Research, 5,* 359-365.

Seeman, J., Barry, E., & Ellinwood, C. (1964). Interpersonal assessment of play therapy outcomes. *Psychotherapy: Theory, Research and Practice, 1,* 64-66.

Shore, M. F., & Massimo, J. L. (1966). Comprehensive vocationally oriented psychotherapy for adolescent delinquent boys: A follow-up study. *American Journal of Orthopsychiatry, 36,* 769-615.

Shore, M. F., & Massimo, J. L. (1990). Five years later: A follow-up study of comprehensive vocationally oriented psychotherapy. *American Journal of Orthopsychiatry, 39,* 769-773.

Shore, M. F., & Massimo, J. L. (1973). After 18 years: A follow-up study of comprehensive vocationally oriented psychotherapy. *American Journal of Orthopsychiatry, 43,* 128-132.

Snyder, J. J., & White, M. J. (1979). The use of cognitive self-instruction in the treatment of behaviorally disturbed adolescents. *Behavior Therapy, 10,* 227-235.

Spence, S. H., & Marzillier, J. S. (1981). Social skills training with adolescent male offenders—II. Short-term, long-term and generalized effects. *Behavioral Research and Therapy, 19,* 349-368.

Stuart, R. B., Tripodi, T., Jayaratne, S., & Camburn, D. (1976). An experiment in social engineering in serving the families of pre-delinquents. *Journal of Abnormal Child Psychology, 4,* 243-261.

Taylor, A. J. W. (1967). An evaluation of group psychotherapy in a girls' borstal. *International Journal of Group Psychotherapy, 17,* 168-177.

Thombs, M. R., & Muro, J. J. (1973). Group counseling and the sociometric status of second grade children. *Elementary School Guidance and Counseling, 7,* 194-197.

Tosi, D. J., Upshaw, K., Lande, A., & Waldron, M. A. (1971). Group counseling with nonverbalizing elementary students: Differential effects of Premack and social reinforcement techniques. *Journal of Counseling Psychology, 18,* 437-440.

Truax, C. B., Wargo, D. G., & Silber, L. D. (1966). Effects of group psychotherapy with high accurate empathy and nonpossessive warmth upon female institutionalized delinquents. *Journal of Abnormal Child Psychology, 71,* 267-274.

Tyler, F. B., & Gatz, M. (1977). Development of individual psychosocial competence in a high school setting. *Journal of Consulting and Clinical Psychology, 45,* 441-449.

Walter, H. I., & Gilmore, S. K. (1973). Placebo vs. social learning effects in parent training procedures designed to alter the behavior of aggressive boys. *Behavior Therapy, 4,* 361-377.

Warner, R. W., Jr. (1971). Alienated students: Six months after receiving behavioral group counseling. *Journal of Counseling Psychology, 18,* 426-430.

Warner, R. W., Jr., & Hansen, J. C. (1978). Verbal-reinforcement and model-reinforcement group counseling with alienated students. *Journal of Counseling Psychology, 17,* 168-172.

Webster-Stratton, C. (1984). Randomized trial of two parent-training programs for families with conduct-disordered children. *Journal of Consulting and Clinical Psychology, 52,* 666-678.

Weinrott, M. R., Corson, J. A., & Wilchesky, M. (1979). Teacher-mediated treatment of social withdrawal. *Behavior Therapy, 10,* 281-294.

White, W. C., Jr., & Davis, M. T. (1974). Vicarious extinction of phobic behavior in early childhood. *Journal of Abnormal Child Psychology, 2,* 25-32.

REFERENCES

Appelbaum, M. I., & Cramer, E. M. (1974). Some problems in the nonorthogonal analysis of variance. *Psychological Bulletin, 81,* 335-343.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasiexperimental designs for research.* Boston: Houghton-Mifflin.

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

Cohen, L. H. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

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

Gould, M. S., Shaffer, D., & Kaplan, D. (1985). The characteristics of dropouts from a child psychiatry clinic. *Journal of the American Academy of Child Psychiatry, 24,* 316-328.

Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart, & Winston.

Hedges, L. V. (1982). Estimation of effect size from a series of independent estimates. *Psychological Bulletin, 92,* 490-499.

Howell, D. C. (1982). *Statistical methods for psychology.* Boston: Duxbury Press.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills: Sage Publications.

Kazdin, A. E. (1978a). Evaluating the generality of findings in analogue therapy research. *Journal of Consulting and Clinical Psychology, 46,* 673-686.

Kazdin, A. E. (1978b). *Research design in clinical psychology.* New York: Harper & Row.

Kraemer, H. C., & Andrews, G. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin, 91,* 404-412.

Lambert, J. M., Shapiro, D. A., & Bergin, A. E. (1986). The effectiveness of psychotherapy. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (3rd ed.). New York: Wiley.

McAdoo, W. G., & Roeske, N. A. (1973). A comparison of defectors and continuers in a child guidance clinic. *Journal of Consulting and Clinical Psychology, 40,* 328-334.

Mintz, J. (1983). Integrating research evidence: A commentary on meta-analysis. *Journal of Consulting and Clinical Psychology, 46,* 71-75.

Rachman, S., & Wilson, G. T. (1980). *The effects of psychological therapy.* Oxford: Pergamon Press.

Shapiro, D. A. (1985). Recent applications of meta-analysis in clinical research. *Clinical Psychology Review, 5,* 13-34.

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

Shapiro, D. A., & Shapiro, D. (1983). Comparative therapy outcome research: Methodological implications of meta-analysis. *Journal of Consulting and Clinical Psychology, 51,* 42-53.

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

Strube, M. J., & Hartmann, D. P. (1983). Meta-analysis: Techniques, applications, and functions. *Journal of Consulting and Clinical Psychology, 51,* 14-27.

Weisz, J. R., & Weiss, B. (1989a). Assessing the effects of clinic-based psychotherapy with children and adolescents. *Journal of Consulting and Clinical Psychology, 57,* 741-746.

Weisz, J. R., & Weiss, B. (1989b). Cognitive mediators of the outcome of psychotherapy with children. In B. B. Lahey & A. E. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 12, pp. 27-51). New York: Academic Press.

Weisz, J. R., Weiss, B., Alicke, M. D., & Klotz, M. L. (1987a). 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., & Langmeyer, D. (1987b). Giving up on child psychotherapy: Who drops out? *Journal of Consulting and Clinical Psychology, 55,* 916-918.

Wilson, G. T., & Rachman, S. J. (1983). Meta-analysis and the evaluation of psychotherapy outcome: Limitations and liabilities. *Journal of Consulting and Clinical Psychology, 51,* 54-64.