Journal of Consulting and Clinical Psychology
1987, Vol. 55, No. 4. 542-549

0022-006X/87/$00.75

Effectiveness of Psychotherapy With Children and Adolescents: A Meta-Analysis for Clinicians

John R. Weisz and Bahr Weiss
University of North Carolina at Chapel Hill

Mark D. Alicke
University of Florida at Gainesville

M. L. Klotz
University of North Carolina at Chapel Hill

How effective is psychotherapy with children and adolescents? The question was addressed by meta-analysis of 108 well-designed outcome studies with 4–18-year-old participants. Across various outcome measures, the average treated youngster was better adjusted after treatment than 79% of those not treated. Therapy proved more effective for children than for adolescents, particularly when the therapists were paraprofessionals (e.g., parents, teachers) or graduate students. Professionals (with doctor's or master's degrees) were especially effective in treating overcontrolled problems (e.g., phobias, shyness) but were not more effective than other therapists in treating undercontrolled problems (e.g., aggression, impulsivity). Behavioral treatments proved more effective than nonbehavioral treatments regardless of client age, therapist experience, or treated problem. Overall, the findings revealed significant, durable effects of treatment that differed somewhat with client age and treatment method but were reliably greater than zero for most groups, most problems, and most methods.

In the late 1970s and early 1980s, experts raised diverse concerns about child and adolescent psychotherapy research. Some complained of insufficient information about therapy outcomes (e.g., Achenbach, 1982). Others suggested that outcome studies revealed few or no effects of therapy (e.g., Gelfand, Jenson, & Drew, 1982). Still others (e.g., Barrett, Hampe, & Miller, 1978) argued that researchers were too preoccupied with the global question of psychotherapy effects per se and should instead study outcomes as a function of treatment approach, type of child, and therapist characteristics. In recent years, the prospects for addressing these concerns have improved considerably. With the development of meta-analytic techniques (Smith, Glass, & Miller, 1980), it is now possible to aggregate findings across multiple studies and to systematically compare findings across dimensions such as treatment approach and client characteristics. The basis for analysis is the effect size, which is computed separately for the treatment group versus control group comparisons of interest. The effect size is an estimate of the magnitude of the treatment effect (treatment group versus control group scores on measures of psychological functioning) adjusted for sample variability.

Meta-analyses of adult outcome studies (Shapiro & Shapiro, 1982; Smith & Glass, 1977) were recently complemented by Casey and Berman's (1985) meta-analysis of studies with children aged 12 years and younger. In the 64 studies that included treatment–control comparisons, the average effect size was .71, indicating a reliable advantage for treatment over no treatment. Although the percentage of boys in the samples was negatively correlated with outcome, Casey and Berman found no substantial differences as a function of age or of group versus individual therapy. Even initial findings showing the superiority of behavioral over nonbehavioral treatments were judged to be artifactual, the result of confounding type of treatment with outcome measure characteristics. The findings suggested that child therapy is demonstrably effective (and about equally so) across age groups and types of therapy. However, before we can conclude that age level is unrelated to therapy effects, we must sample adolescents as well as children. Moreover, as Parloff (1984) noted, findings suggesting that different therapies work equally well deserve close scrutiny because their implications are so significant.

Might therapy effects be different for adolescents than for children? Various cognitive and social developments (see Rice, 1984), including the advent of formal operations (see Piaget, 1970), make adolescents more cognitively complex than children, less likely to rely on adult authority, and seemingly less likely to adjust their behavior to fit societal expectations (see Kendall, Lerner, & Craighead, 1984). Thus, it is possible that adolescents may be more resistant than children to therapeutic intervention. On the other hand, adolescents are more likely than children to comprehend the purpose of therapy and to understand complex, interactive psychological determinants of behavior, which could make them better candidates for therapy than children. Because most child clinicians actually treat both children and adolescents, a comparison of therapy effects in

This project was facilitated by support from the North Carolina Division of Mental Health, Mental Retardation, and Substance Abuse Services (Department of Human Resources) and by National Institute of Mental Health Grant R01 MH 38240-01. We are grateful to David Langmeyer for his support and suggestions, to Gary Bornstein, Larry Crum, and Lynn Fisher for their assistance in data collection and computation, and to Thomas Achenbach for his thoughtful review of an earlier draft of this article.

Correspondence concerning this article should be addressed to John R. Weisz, Department of Psychology, Davie Hall 013A, University of North Carolina, Chapel Hill, North Carolina 27514.

these two age groups could have implications for their work. To effect this comparison, we reviewed studies of youngsters aged 4–18 years.

We also reexamined the question of whether therapy methods differ in effectiveness. Casey and Berman (1985) found that an apparent superiority of behavioral over nonbehavioral methods largely evaporated when they excluded cases in which outcome measures "were very similar to activities occurring during treatment" (p. 391). Our own review suggested that many of the comparisons that were dropped may actually have involved sound measurement of therapy effects. Consider, for example, interventions in which phobic children are exposed to models who bravely approach animals that the phobic children fear. The most appropriate outcome assessment may well be one that is similar to therapy activities–behavioral tests of the children's ability to bravely approach the animals. Certainly, such tests may be less than ideal; for example, they may involve contrived circumstances that differ from the clinical situation of primary interest. However, in many such tests, children who perform the target behavior may well be displaying not a measurement artifact but real adaptive behavior. On the other hand, some of the Casey and Berman (1985) exclusions seem quite appropriate. For example, researchers who use the Matching Familiar Figures Test (MFFT) to teach reflectivity should not use the MFFT to assess outcomes because the ultimate goal is broader than improved MFFT performance. In this case, the similarity between treatment activity and outcome measure is inappropriate because it is unnecessary.

To further examine the issue of behavioral–nonbehavioral differences, we distinguished between two kinds of comparisons in which outcome measures were similar to training activities: (a) comparisons in which such similarity was necessary for a fair and valid assessment and (b) comparisons in which such similarity was unnecessary and posed a risk of artifactual findings. In a key behavioral–nonbehavioral contrast, we included comparisons in the first group but not those in the second.

Our third objective was to examine treatment effects as a function of treated problem. However, we focused not only on the specific problem categories used in other meta-analyses (e.g., phobias, impulsivity) but also on the two overarching, broadband categories most often identified in factor analytic research (Achenbach, 1982; Achenbach & Edelbrock, 1978): overcontrolled (e.g., phobias, shyness) and undercontrolled (e.g., aggression, impulsivity).

Finally, we questioned whether therapy effects differ with level of therapist training. The evidence thus far is mixed but is often discouraging for trained therapists in general (see Auerbach & Johnson, 1977; Parloff, Waskow, & Wolfe, 1978). However, there is little summary evidence on therapists who work with young clients in particular, which is an unfortunate gap given the substantial time and resources invested in the professional training of child clinicians. We attempted to fill this gap. Recognizing that the effects of training might differ depending on client age, on the treatment method used, or on the problem being treated, we also explored interactions of training with each of these factors.

Method

Defining Psychotherapy

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 a predetermined treatment plan. We excluded approaches involving drug administration, reading material only (bibliotherapy), teaching or tutoring only to increase knowledge of a specific subject, moving youngsters to a new living situation (e.g., a foster home), and efforts solely to prevent problems to which youngsters were deemed at risk.

We did not require that psychotherapy be conducted by fully trained professionals. Some schools of thought hold that extensive professional training is not required for effective interventions but that parents, teachers, or siblings may function as effective change agents. This suggested to us that, rather than prejudge the issue, we should treat the impact of therapist training as an empirical question in the analysis.

Literature Search

Several approaches were combined to identify relevant published studies. A computer search was carried out using 21 psychotherapy-related key words and synonyms, and the resulting data were crossed with appropriate age group and topic constraints.[1] This helped us to rule out the myriad of articles that merely described or advocated certain therapy methods. The initial pool of 1,324 articles was reduced in stepwise fashion by using title, abstract, and method-section information to select only those studies that met our inclusion requirements. We used three other procedures to enhance the comprehensiveness of our survey. First, all articles cited in the meta-analyses by Smith, Glass, and Miller (1980) and by Casey and Berman (1985) were surveyed and included if they met our selection criteria. Second, *Psychological Abstracts* entries from January 1970 to September 1985 were searched by hand. Third, four journals accounting for the majority of appropriate articles in the preceding steps were searched, issue by issue, for the same time period. These were *Behavior Therapy,* the *Journal of Abnormal Psychology,* the *Journal of Consulting and Clinical Psychology,* and the *Journal of Counseling Psychology.* The result of the search was a pool of 108 controlled studies of psychotherapy outcomes among children and adolescents that met our criteria.[2] The psychological literature accounted for more of the studies than the psychiatry, social work, or nursing literature. Of the 108 studies, 24 (22%) were also included in the Smith et al. (1980) analysis and 32 (30%) were also included in the Casey and Berman (1985) analysis.

Subject Population

The studies focused on the prekindergarten through secondary school age range (i.e., 4–18 years). Across the 108 studies, mean age was 10.23 (SD = 4.00). Approximately 66% of the youngsters sampled were male.

---

[1] The 21 psychotherapy key words and synonyms were client-centered, contract- (ing, systems, etc.), counseling, cotherapy, dream analysis, insight-, intervention-, model-, modifica-, operant-, paradox-, psychoanaly-, psychodrama-, psychothera-, reinforce-, respondent, role-playing, therap-, training, transactional, and treatment. The age group constraints were adolescen-, child-, juvenile-, pre-adolescen-, and youth-. The evaluation-oriented topic constraints were assess-, comparison, effect-, efficacy, evaluat-, influence, impact, and outcome.

[2] A list of the studies included in this meta-analysis is available from the first author for a $5 fee to cover printing, postage, and handling.

To be consistent with other reviews (e.g., Casey & Berman, 1985; Smith & Glass, 1977), we included studies focusing on a broad variety of psychological or adjustment problems. However, we excluded mental retardation; underdeveloped reading, writing, or knowledge of specific school subjects; problems involving seizures; and physically disabling handicaps. Problems that have been attributed to physiological causes but for which etiology has not been well-established (e.g., attention deficit, hyperactivity, and learning disability) were included provided that a behavioral or psychological problem (e.g., impulsivity) was actually addressed in treatment.

Design and Reporting Requirements

We required that a study compare a treated group with an untreated or minimally treated control group and that the control condition provide little more than attention to the youngsters. We classified control groups that had provided alternate treatment or one element of a full treatment package as treatment groups. If such treatment constituted the only control condition included, the study was dropped. We also excluded studies that used subjects as their own controls in single-subject or within-subject designs. Such studies generate an unusual form of effect size (e.g., based on intrasubject variance, which is not comparable to conventional variance statistics) and do not appear to warrant equal weighting with studies that include independent treatment and control samples.

Classification and Coding Systems

Studies were coded for sample, treatment, and design characteristics, with some coding systems patterned after those of Casey and Berman (1985).[3] Coding and effect size calculation were carried out independently to avoid bias. One fourth of the studies were randomly selected for independent coding by two judges.

Treatment approaches. Treatment methods were classified using the three-tiered system shown in Table 1. Tier 1 included the broad categories of behavioral and nonbehavioral. Tier 2 included the subcategories (e.g, respondent procedures) grouped within each Tier 1 category. Tier 3 included fine-grained descriptors (e.g., extinction); only the behavioral studies could be classified this finely. Despite what were often quite limited descriptions of treatment methods, the two raters achieved kappas of .74, .71, and .78 on Tiers 1, 2, and 3, respectively. Two of the 163 comparisons (1%) were described too vaguely to be coded. We also coded treatment approaches as group-administered or individually-administered (κ = .92).

Target problem. Treated problems were coded using the two-tiered system shown in Table 3. At the most general level, problems were grouped into the two broadband categories most often identified in factor analyses of child and adolescent behavior problems: undercontrolled (e.g., aggressive, acting out, or externalizing behavior) and overcontrolled (e.g., shy or withdrawn, phobic, or internalizing behavior; see Achenbach & Edelbrock, 1978). Problems not fitting either category were coded as *other.* The second tier consisted of descriptive subcategories (e.g., shy, withdrawn; phobias, specific fears). The two raters achieved kappas of .94 and .86 for Tiers 1 and 2, respectively. In 9 of the 163 group comparisons (6%), the problems were described too vaguely to be coded.

Outcome measures. We used Casey and Berman's (1985) system to code whether outcome measures were similar to treatment activities (our κ = .82). As noted previously, we also carried out one further coding of outcome measures that were rated similar to treatment activities: We coded for whether the similarity was necessary for a fair assessment (given the treatment goals) or unnecessary (κ = .81). We also used Casey and Berman's (1985) category systems for coding outcome measures into type and source. Seven types were included, and five types occurred with sufficient frequency to warrant kappa calculations. The categories were (a) fear/anxiety (κ = .98); (b) cognitive skills (κ = .87); (c) global adjustment (κ = .83); (d) social adjustment (κ = .91); (e) achievement (including school grades and achievement test scores; κ = .98) (f) personality (including scales measuring attitudes and beliefs); and (g) self-

Table 1
Mean Effect Size for Each Therapy Type

Therapy type	No. treatment groups	Effect size	p[a]
Behavioral[b]			
Operant	39	.78	.0001
Physical reinforcers	8	.92	.07
Consultation in operant methods	16	.77	.0001
Social/verbal reinforcement	6	.78	.002
Self-reinforcement	2	.33	.40
Combined physical and verbal reinforcement	3	.75	.01
Multiple operant methods	2	.72	.07
Respondent	17	.75	.0002
Systematic desensitization	8	.65	.01
Relaxation (no hierarchy)	3	.43	.05
Extinction (no relaxation)	3	1.46	.14
Combined respondent	2	.43	.39
Modeling	25	1.19	.003
O'Connor film[c]	4	2.90	.27
Live peer model	6	1.25	.02
Live nonpeer model	3	.62	.26
Nonlive peer model	9	.79	.0001
Nonlive nonpeer model	2	.29	.21
Social skills training	5	.90	.04
Cognitive/cognitive behavioral	10	.68	.0004
Multiple behavioral	10	1.04	.0002
Nonbehavioral[d]			
Client-centered/nondirective	20	.56	.0001
Insight-oriented/psychodynamic	3	.01	.98
Discussion group	4	.18	.18

Note. Because some descriptions of treatment methods were too vague to be classified, not all *N*s sum as expected.

[a] The probability that a particular effect size is greater than zero reflects the variability of effect sizes across the category of comparisons being sampled in addition to the number of studies and the mean effect size. Thus, where effect sizes are quite variable across a category, the p value may be low despite a high mean effect size and a substantial pool of studies. A Bonferroni correction applied to the tests in this table indicated that probability values $\geq$ .002 should not be regarded as statistically significant.

[b] N = 126; effect size = .88; $p < .0001$.

[c] The O'Connor film shows a child engaging in social entry behavior. One study using this film reported an unusually high effect size (9.5); with this study dropped, effect size = .77.

[d] N = 28; effect size = .42; $p < .0001$.

[3] We used Casey and Berman's (1985) coding schemes for (a) whether outcome measures were similar to activities occurring during treatment, (b) type of outcome measure, and (c) source of outcome measure. In addition, we distinguished between behavioral and nonbehavioral therapies, as did Casey and Berman, but we also distinguished among subtypes within each broad category.

Table 2
Mean Effect Size for Behavioral and Nonbehavioral Treatments

Analysis	Behavioral		Nonbehavioral			
	M	*N*	*M*	*N*	*t*	*p*
All comparisons	.88	126	.44	27	2.14	< .05
Omitting therapy-like outcomes	.61	34	.51	22	.64	.52
Omitting unnecessary therapy-like outcomes	.93	94	.45	24	2.09	< .05

Note. All six effect size means were significantly different from zero (all *ps* < .0006).

concept. The kappa calculation for the full system was .90. Casey and Berman coded eight sources of outcome measures in their system: (a) observers (κ = .90); (b) therapists; (c) parents (κ = .96); (d) subject's own performance (κ = .91); (e) expert judges; (f) peers (κ = .90); (g) teachers (κ = .97); and (h) self-report by subjects (κ = .96). The kappa calculation for the full system was .92.

Therapist training. We classified therapists according to level of training. Levels included (a) professionals who held a doctor's or master's degree in psychology, education, or social work; (b) graduate students who were working toward advanced degrees in psychology, education, or social work; and (c) paraprofessionals who were parents, teachers, or others lacking mental-health-related graduate training but trained to administer the therapy.

Calculation of Effect Sizes

Effect sizes were estimated using the procedures of Smith et al. (1980, Appendix 7). In each calculation, the mean posttherapy treatment group–control group difference was divided by the control group standard deviation.[4] Most studies in our pool included multiple outcome measures, and a number of studies included more than one treatment condition. Thus, in most cases, a study initially produced numerous possible effect sizes. To retain all effect sizes in our analysis would have resulted in disproportionate weighting of those studies with the most measures and groups. Several solutions to this problem are available (e.g., Glass, McGaw, & Smith, 1981). We chose to collapse effect sizes across outcome measures except in analyses comparing such measures. However, the comparison of treatment procedures was central to the overall analysis, and separate treatments within studies appeared sufficiently independent to warrant separate attention. Consequently, for each study we computed one effect size estimate for each of the treatment conditions included. The 108 studies included an average of 1.54 therapy conditions for a total of 163 effect sizes over the entire pool. Some studies included both follow-up and posttreatment assessments of therapy effects. We describe follow-up findings separately in a final section.

Results

Overview of Procedures for Analysis

Two problems often present in meta-analyses are that most group comparisons involve tests of main effects alone and that key variables (e.g., treatment method and treated problem) are confounded (see Glass & Kliegl, 1983; Mintz, 1983). The few cases that have attempted more controlled comparisons have often dropped substantial portions of available data. Here, we selected an approach intended to use all the data available while avoiding undue risk of either Type I or Type II error. Minimizing Type II error is particularly important in meta-analyses given its potential heuristic, hypothesis-generating value.

The first wave of analysis focused planned comparisons on our four factors of primary interest: age group, therapy type, target problem type, and therapist training. For these analyses, we tested the four main effects. We then tested each main effect for its robustness, using (a) general linear models (GLM) procedures that eliminated (i.e., controlled for) the effects of each of the other three factors (see Appelbaum & Cramer, 1974) and (b) tests of whether any of the main effects were qualified by interactions with any of the other three factors.[5] We also tested the robustness of the therapy type effect using the Casey and Berman (1985) procedure followed by our own revised procedure. For all other group comparisons, we applied a Bonferroni correction (Neter & Wasserman, 1974), which set the alpha at .01. Bonferroni corrections were also applied to each family of tests to compare obtained effect sizes to the null hypothesis of zero (see Tables 1, 3, and 4).

Overall Effect Size

Across the 163 treatment–control comparisons, the mean effect size was 0.79 (significantly different from zero, $p < .0001$). The average treated youngster was placed at the 79th percentile of those not treated. Of the 163 effect sizes, only 10 (6%) were negative, indicating an adverse effect of treatment. The mean effect size of 0.79 was comparable to those reported in earlier meta-analyses of therapy effects among children (0.71 in Casey & Berman, 1985), adults (0.93 in Shapiro & Shapiro, 1982), and mixed age groups (0.68 in Smith & Glass, 1977).

Preliminary Check for Sex Effects

Before proceeding to the group comparisons of primary interest, we checked the relation between effect size and gender composition of treated groups. For 99 of the treatment–control comparisons, information was sufficient to reveal whether a majority of the treatment group comprised male (N = 72) or female (N = 27) participants. Effect sizes averaged 0.80 for the male majority groups and 1.11 for the female majority groups (p value for difference = .33). Studies that did not report gender composition averaged an effect size of 0.55.

[4] Some researchers (e.g., Casey & Berman, 1985; Hedges, 1982) favor the use of a pooled treatment and control group standard deviation as denominator. If one consequence of therapy is an increase in behavioral variability, as some researchers have suggested (e.g., Bergin & Lambert, 1978), such pooling can cause interpretational and statistical problems (see Smith et al., 1980), which we sought to avoid.

[5] We considered using full factorial model ANOVAS, but cell sample sizes were too unbalanced to yield interpretable results and some were so small that tests of 3-way and 4-way interactions were impossible.

Age Level

A question of primary interest was whether the impact of therapy differed for children and adolescents. The mean effect size for the 98 treatment–control comparisons involving children (ages 4–12 years) was 0.92 (82nd percentile), which was significantly larger than the mean of 0.58 (72nd percentile) for the 61 comparisons involving adolescents (ages 13–18 years), $t(157) = 2.17$, $p < .05$. The correlation between age and effect size was −0.21 ($p < .01$) over the entire sample; the coefficient was −0.17 ($p < .10$) for effect sizes involving children and 0.15 (*ns*) for those involving adolescents.

We next tested the robustness of the age group differences by using eliminating tests. The age effect was reduced slightly when therapy type (behavioral vs. nonbehavioral) was controlled ($p = .084$) and when problem type (overcontrolled vs. undercontrolled) was controlled ($p = .086$). Both reductions were caused partly by reduced sample size because not all treatments or target problems could be coded. However, the age group difference grew more reliable statistically when therapist training was controlled ($p = .013$).

A series of 2 × 2 analyses of variance (ANOVAs) testing interactions of age with therapy type, problem type, and therapist training, respectively, revealed no significant effects (all Fs < 1.8; all ps > .15). To provide the most thorough assessment, we carried out GLM ANOVAs on the same interactions, with age entered as a continuous variable. Neither therapy type nor problem type interactions were significant (both ps > .30), but the Age × Therapist Training interaction was significant, $F(2, 102) = 3.49$, $p < .05$. Age and effect size were uncorrelated among professionals ($N = 39$, $r = 0.11$, $p < .50$) but were negatively correlated among graduate students ($N = 43$, $r = -0.31$, $p < .05$) and paraprofessionals ($N = 26$, $r = -0.43$, $p < .05$). Trained professionals were about equally effective with all ages, but graduate students and paraprofessionals were more effective with younger than with older clients.

Therapy Type

Behavioral versus nonbehavioral. Table 1 shows that mean effect size was higher for behavioral than for nonbehavioral treatments, $t(152) = 2.14$, $p < .05$. The difference remained significant after eliminating tests controlled for age ($p < .05$), problem type ($p < .05$), and therapist training ($p < .05$). Interactions of therapy type with problem type and with therapist training were not significant (both Fs < 0.50, both ps > .65).

Similarity of therapy procedures and outcome measures. Next, we examined the main effect for therapy type, following-up on Casey and Berman's (1985) analysis. We first excluded comparisons involving an outcome measure similar to treatment procedures. Consistent with Casey and Berman's (1985) finding, the procedure reduced the behavioral–nonbehavioral difference to nonsignificance, $t(55) = 0.64$, $p = .52$. As we noted earlier, the Casey and Berman procedure may rule out some carefully designed studies in which measures similar to the training procedures are appropriate and necessary for a fair test of treatment success. To correct for this apparent limitation, we again compared behavioral and nonbehavioral methods and restored to the sample all treatment–control comparisons in which similarity of training and assessment methods was judged by our raters to be necessary for a fair test. In this analysis, as Table 2 shows, behavioral-method comparisons showed a significantly larger effect size than nonbehavioral comparisons, $t(117) = 2.09$, $p < .05$.

Table 3
Mean Effect Size for Each Target Problem

Target problem	No. treatment groups	Effect size	p^a
Undercontrolled	76	.79	.0001
Delinquency	19	.66	.0004
Noncompliance	9	1.33	.005
Self-control (hyperactivity, impulsivity)	31	.75	.0001
Aggressive/undisciplined	17	.74	.0002
Overcontrolled	67	.88	.0001
Phobias/anxiety	39	.74	.0001
Social withdrawal/isolation	28	1.07	.002
Other	18	.56	.0002
Adjustment/emotional disturbance	9	.69	.001
Underachievement	9	.43	.046

Note. Because some problem descriptions were too vague to be classified, not all *N*s sum as expected.
[a] A Bonferroni correction applied to this table indicated that probability values ≥ .005 should not be regarded as statistically significant.

Specific therapy types. The ANOVAs focused on Tier 2 of the therapy-type coding system (see Table 1) revealed no significant difference between the behavioral subtypes (e.g., operant, modeling) or the nonbehavioral subtypes (e.g., psychodynamic, client centered), all Fs < 1.9; all ps > .15. The comparison of nonbehavioral subtypes should be interpreted with caution: The majority of these studies used client-centered therapy and only three used insight-oriented psychodynamic therapy. The ANOVAs focused on Tier 3 revealed no significant differences between the therapies within each subtype. As Table 1 shows, effect sizes for most of the categories within each tier were significantly greater than zero.

Target Problem

Overcontrolled versus undercontrolled problems. Next, we focused on the problems for which the youngsters were treated. There was no significant difference between the broad categories of overcontrolled and undercontrolled problems ($p = .46$; see Table 3). This continued to be true when eliminating tests were used to control for age level ($p = .59$), therapy type ($p = .67$), and therapist training ($p = .67$).

However, problem type and therapist training did interact, $F(2, 90) = 2.93$, $p = .059$. Tests of simple effects revealed no significant differences in effect size between overcontrolled and undercontrolled problems at any of the three levels of therapist training. Nor did the three therapist groups differ in their success with undercontrolled youngsters. The groups did differ, however, in their effectiveness with overcontrolled youngsters: As amount of formal training increased, so did effectiveness,

Table 4
Effect Size as a Function of Source of Outcome Measure

Source	*N*	Effect size	*p*[a]
Observers	62	1.08	.0001
Parents	19	.66	.0001
Teachers	48	0.68	.0001
Subject performance	128	0.65	.0001
Subject report	61	0.49	.0001
Peers	18	0.33	.01

[a] A Bonferroni correction applied to this table indicated that probability values > .008 should not be regarded as statistically significant.

linear trend $F(1, 35) = 4.68$, $p < .05$. Professionals achieved a mean effect size of 1.03, graduate students of 0.71, and paraprofessionals of 0.53.

Specific problem types. The categories in Tier 2 of the problem coding scheme (e.g., delinquency, specific fears) did not differ significantly in effect size ($p = .41$). However, as Table 3 shows, effect sizes were reliably greater than zero for most problem categories.

Therapist Training

Although therapist training entered into two interactions, its main effect was not significant ($p = .43$). This continued to be true when eliminating tests were used to control for age ($p = .56$), therapy type ($p = .65$), and problem type ($p = .51$).

Other Findings

Individual versus group therapy. Does it matter whether youngsters are treated individually or in groups? Our data revealed somewhat larger effect sizes for therapy that was individually administered rather than group administered ($M = 1.04$ and $M = 0.62$, respectively), but the difference did not attain significance under our Bonferonni correction procedure ($p = .03$).

Source and content of outcome measure. Casey and Berman (1985) found significant differences as a function of the source of outcome measures (e.g., observers, teachers). Measures derived from observers revealed the largest difference between the treated and control groups. Using their category system (with two low-frequency categories dropped), we also found a main effect for source, $F(5, 330) = 4.00$, $p < .005$. Newman-Keuls tests indicated that observers reported more change than any of the other sources (all $ps < .05$), none of which differed significantly from one another (see Table 4).

Casey and Berman also reported significant differences as a function of the content of outcome measures (e.g., fear and anxiety, cognitive skills). We used their category system but failed to find a significant main effect ($p = .61$). We dropped two low frequency categories, but the main effect remained nonsignificant ($p = .42$).

Clinic-referred versus analog samples. Outcome research, and meta-analyses of such research, have been criticized for excessive reliance on *analog* samples, that is, samples that have been recruited by researchers specifically for treatment studies rather than samples that have been spontaneously referred by clinics (e.g., Parloff, 1984; Shapiro & Shapiro, 1982). Combining outcome results from analog and clinic samples makes it difficult to gauge the relevance of findings to actual clinic practice. For this reason, we separated the 126 comparisons involving analog samples from the 37 comparisons involving true clinical samples. Mean effect sizes were 0.76 for analog samples and 0.89 for clinical samples: The difference was not significant ($F < 1$).

Follow-up findings: Do therapy effects last? The preceding findings suggest that therapy does have positive effects that are measurable at the end of therapy. To have real practical value, however, the effects must be durable and must persist beyond treatment termination. To assess the durability of therapy effects, we analyzed the follow-up treatment–control comparisons contained in our sample; these follow-ups averaged 168 days subsequent to termination of treatment. Average effect size at follow-up (0.93) was actually larger than effect size immediately after treatment (0.79), although the difference was not significant ($p = .45$). When we included in the posttreatment group only those studies that also had a follow-up assessment, the means for posttreatment and follow-up were identical (0.93). Thus, the effects of therapy in this pool of studies appear to be durable.

Discussion

Is psychotherapy effective with children and adolescents? The findings reviewed here suggest that it is. After treatment, across multiple measures of adjustment, the average treated youngster was functioning better than 79% of those not treated. However, the benefits of therapy depended to some extent on the age level of the youngsters treated, with children profiting more than adolescents. This is consistent with the idea, suggested earlier, that cognitive changes such as the advent of formal operations (Piaget, 1970), and other quasicognitive changes (Perlmutter, 1986; Rice, 1984), may make adolescents less responsive than children to therapeutic influence; enhanced powers of reasoning may strengthen adolescents' convictions regarding their own behavior and may also make them adept at circumventing or sabotaging a therapist's efforts. Of course, it is also possible that outcome measures used with adolescents are less sensitive to change than most child measures or that adolescents' problems are more entrenched (see Kendall et al., 1984, for further ideas about age group differences and their sources).

Only paraprofessionals and graduate student therapists were more effective with younger than older children; trained professionals were about equally effective with younger and older clients. Taken together, the age main effect and the Age × Training interaction suggest an intriguing possibility. It may be that adolescents, for cognitive or other reasons, are generally more difficult than children to treat successfully, but formal training may provide professionals with sufficient therapeutic acumen to override age differences in initial treatability.

Our findings on therapy type do not support the often noted summary of adult psychotherapy findings that different forms

of therapy work about equally well (for a critical discussion, see Parloff, 1984; see also Frank's, 1973, "nonspecificity hypothesis"). Instead, we found that behavioral methods yielded significantly larger effects than nonbehavioral methods. This finding held up across differences in age level, treated problem, and therapist experience, and it was not qualified by interactions with any of these factors. The behavioral–nonbehavioral difference was reduced to nonsignificance when we excluded all therapy-like outcome measures (following Casey & Berman, 1985) but was readily revived when we excluded only the unnecessary therapy-like measures that seemed likely to produce artifactual findings. Overall, the findings make a case for the superiority of behavioral over nonbehavioral approaches.

On the other hand, we found comparatively few controlled studies assessing nonbehavioral methods; it might be argued that these studies do not represent the best of nonbehavioral approaches. In fact, many would argue that nonbehavioral approaches are best suited to the only partly scrutable process of unraveling complex causal dynamics and of stimulating insight over months or even years of treatment. By contrast, most controlled-outcome research focuses on relatively specific target problems, directly observable outcome measures, and brief treatment. Do such studies miss the point of nonbehavioral intervention? Perhaps, but the present findings seem to place the burden of proof on those who make that argument.

Those familiar with the evidence on the long-term stability of undercontrolled behavior and the relative instability of overcontrolled behavior (reviewed in Robins, 1979) may have been surprised to find that therapy effects were no more pronounced with the latter than the former. Note, though, that the evidence on long-term stability concerns the tendency of problems to persist or dissipate over time, independent of therapy. Our review, by contrast, focuses on the persistence of various problems in treated groups relative to control groups. When the natural dissipation of problems over time is thus held constant, our findings suggest that undercontrolled problems may be no more intractable than overcontrolled problems.

Our failure to find an overall difference in effectiveness between professionals, graduate students, and paraprofessionals might be disquieting to those involved in clinical training programs (see also Auerbach & Johnson, 1977; Parloff, Waskow, & Wolfe, 1978). A simplistic interpretation might suggest that training does not enhance therapeutic effectiveness. A more thoughtful evaluation, though, suggests otherwise. First, it should be noted that the therapeutic work of the graduate students and paraprofessionals did not take place in a vacuum: In nearly every instance, these therapists were selected, trained, and supervised by professionals in techniques that professionals had designed. Thus, the success enjoyed by the two less clinically trained groups might actually have reflected the judgment and skill of professionals working behind the scenes.

Moreover, the finding of no overall difference was qualified by two important interactions, both suggesting possible benefits of training. A Training × Age interaction suggested that professional training may enhance therapist effectiveness with older, more difficult-to-treat children. And a Training × Problem Type interaction suggested that training may enhance therapist effectiveness in treating overcontrolled problems. On a more negative note, this interaction suggested that training may have little impact on therapists' effectiveness with undercontrolled problems. Why? One possibility is that youngsters with undercontrolled problems are responsive to interventions that are relatively easy to learn; some of these interventions may be similar to natural, naive, parent-like responses that arise in situations requiring discipline and rule enforcement.

Two findings were particularly encouraging. The first revealed that therapy studies with analog samples yielded results quite similar to studies with true clinic-referred samples. Thus, we found no evidence that the positive findings generated by analog therapy studies presented an inflated or otherwise distorted picture of therapy effects among children and adolescents. A second source of encouragement was the finding that the impact of therapy assessed at immediate posttest was not reliably different from the impact assessed at follow-up, which occurred an average of 6 months later. This is consistent with other findings across a broader age range (reviewed by Nicholson & Berman, 1983) indicating that therapy effects may be relatively enduring.

Here, and in most of the other findings, there is reason for optimism about therapy effects with children and adolescents. On the other hand, the number of available outcome studies is still much too modest to permit a definitive analysis. A well-developed understanding of therapy outcomes within the broad range sampled here will require the continued efforts of our best researchers.

References

Achenbach, T. M. (1982). *Developmental psychopathology* (2nd ed.). New York: Wiley.

Achenbach, T. M., & Edelbrock, C. S. (1978). The classification of child psychopathology: A review and analysis of empirical efforts. *Psychological Bulletin, 85,* 1275–1301.

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

Auerbach, A. H., & Johnson, M. (1977). Research on the therapist's level of experience. In A. S. Gurman & A. M. Razin (Eds.), *Effective psychotherapy: A handbook of research* (pp. 84–99). New York: Pergamon Press.

Barrett, C. L., Hampe, I. E., & Miller, L. (1978). Research on psychotherapy with children. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (pp. 411–435). New York: Wiley.

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. 139–189). New York: Wiley.

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

Frank, J. D. (1973). *Persuasion and healing.* Baltimore, MD: Johns Hopkins University Press.

Gelfand, D. M., Jenson, W. R., & Drew, C. J. (1982). *Understanding child behavior disorders.* New York: Holt, Rinehart, & Winston.

Glass, G. V., & Kliegl, R. M. (1983) An apology for research integration in the study of psychotherapy. *Journal of Consulting and Clinical Psychology, 51,* 28–41.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

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

Kendall, P. C., Lerner, R. M., & Craighead, W. E. (1984). Human development and intervention in childhood psychopathology. *Child Development, 55,* 71–82.

Mintz, J. (1983). Integrating research evidence: A commentary on meta-analysis. *Journal of Consulting and Clinical Psychology, 51,* 71–75.

Neter, J., & Wasserman, W. (1974). *Applied linear statistical models.* Homewood, IL: Irwin.

Nicholson, R. A., & Berman, J. S. (1983). Is follow-up necessary in evaluating psychotherapy? *Psychological Bulletin, 93,* 261–278.

Parloff, M. B. (1984). Psychotherapy research and its incredible credibility crisis. *Clinical Psychology Review, 4,* 95–109.

Parloff, M. B., Waskow, I. E., & Wolfe, B. E. (1978). Research on therapist variables in relation to process and outcome. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (pp. 233–274). New York: Wiley.

Perlmutter, M. (Ed.). (1986). *Cognitive perspectives on children's social and behavioral development: The Minnesota symposia on child psychology* (Vol. 18). Hillsdale, NJ: Erlbaum.

Piaget, J. (1970). Piaget's theory. In P. E. Mussen (Ed.), *Carmichael's manual of child psychology* (3rd ed., Vol. 1, pp. 703–732). New York: Wiley.

Rice, F. P. (1984). *The adolescent: Development, relationships, and culture.* Boston: Allyn and Bacon.

Robins, L. N. (1979). Follow-up studies. In H. C. Quay & J. S. Werry (Eds.), *Psychopathological disorders of childhood* (2nd ed., pp. 483–513). New York: Wiley.

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

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

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

Received July 31, 1986
Revision received December 1, 1986
Accepted December 1, 1986 ■