The Clinical Significance of Treatments: A Comparison of Three Treatments for Conduct Disordered Children

Radley C. Sheldrick, Philip C. Kendall, and Richard G. Heimberg
Temple University

We demonstrate two methods of assessing clinical significance by comparing three treatments for conduct-disordered children. Clinical significance was examined by exploring two questions. First, the question of whether the change attributable to treatment was of a large enough magnitude to be considered clinically significant was examined using the reliable change index. Second, whether treated individuals were distinguishable from normal individuals with regard to target variables was examined using normative comparisons conducted with a statistical technique known as equivalency testing. Three treatments meeting criteria for well-established or probably efficacious treatments were reviewed. All three produced clinically significant changes. However, significant differences were found in terms of normative comparisons.

Key words: **clinical significance, normative comparisons, treatment outcome, psychotherapy evaluation, reliable change index.** ***[Clin Psychol Sci Prac 8:418–430, 2001]***

Randomized clinical trials (RCTs) are designed to determine whether statistically significant change can be attributed to a given treatment. However, RCTs do not address whether a treatment has significant and beneficial impact on participants' lives—that is, whether the results of that treatment are clinically significant. To evaluate clinical significance, distinct measures and methods may be used. Clinical researchers must choose salient variables to investigate, decide how these variables should be measured, employ appropriate statistics to evaluate their results, and determine the optimal method for judging clinical significance. Each of these topics requires careful consideration. This article focuses on judging clinical significance and offers a method by which the clinical significance of treatments may be evaluated and compared.

Over the past two decades, several statistical methods for evaluating clinical significance have been put forward. Among these, two are particularly useful for evaluating and comparing the clinical significance of treatments: the reliable change index (RCI) and normative comparisons. To examine the use of these two methods, we evaluated the clinical significance of three treatments for conduct-disordered children.

Externalizing disorders in children are among the most often diagnosed disorders (Reid, 1993) and among the most stable of childhood problems (Esser, Schmidt, & Woerner, 1990; Pope & Bierman, 1999; Vuchinich, Bank, & Patterson, 1992). Furthermore, early conduct problems place children at risk for future difficulties, including peer rejection and school failure in middle childhood (McMahon, 1994), as well as diagnosable psychopathology (Egeland, Pianta, & Ogawa, 1996), substance abuse, and arrest for criminal behavior in adolescence (Bank, Duncan, Patterson, & Reid, 1994; Capaldi & Clark, 1998; Moffitt, Caspi, Dickson, Silva, & Stanton, 1996; Patterson, Forgatch, Yoerger, & Stoolmiller, 1998; Stattin & Magnusson, 1996).

Treatments for childhood behavior problems are among the most studied of psychological interventions. In their review of effective psychosocial treatments for conduct-disordered children and adolescents, Brestan and Eyberg (1998) identified 82 treatment studies, as well as 12 distinct treatment methods that meet the criteria set forth for either "well-established" or "probably effica-

Address correspondence to Philip C. Kendall, Department of Psychology, 1701 North 13th Street, Weiss Hall, Temple University, Philadelphia, PA 19122.

cious" therapies (Chambless & Hollon, 1998). To qualify for either designation, treatments were required to be standardized in a treatment manual. Furthermore, well-established treatments had to have demonstrated superiority to placebo or an alternative treatment, as well as replication by an independent investigator. Treatments qualified as probably efficacious if at least two studies demonstrated their superiority over waitlist controls or if they met all criteria for well-established therapies except replication by an independent investigator.

These criteria, however, do not specifically address the issue of clinical significance. Questions of clinical significance refer to whether changes affected by treatment are meaningful and/or convincing. Tests of statistical significance used in randomized clinical trials are appropriate to demonstrate a treatment's ability to affect change beyond chance levels, but they do not demonstrate that this change is clinically meaningful (Kendall, Flannery-Schroeder, & Ford, 1999). Thus, the criteria listed, which rely on traditional statistical tests of significance, are not sufficient to demonstrate clinical significance.

In their discussion of criteria for effective treatments, Chambless and Hollon (1998) mentioned the issue of clinical significance in treatment outcome research as one of great importance. However, they did not raise attainment of a particular level of clinical significance to the status of a criterion for "well-established" or "probably efficacious" treatments for two reasons. First, at present, few studies report data on clinical significance. For example, although all of the 82 studies in Brestan and Eyberg's (1998) review were published in peer-reviewed journals and the majority exhibited a high degree of methodological rigor, only a few addressed the issue of clinical significance. Second, no standard, well-accepted guidelines for the evaluation of clinical significance existed that could have been applied to the broad range of studies. Here we demonstrate a method by which these issues may be addressed by examining the clinical outcome literature in the area of conduct-disordered children in light of clinical significance.

ASSESSING CLINICAL SIGNIFICANCE

The issue of whether the results of a given treatment outcome study are clinically significant may by conceptualized by asking two questions: Was the change attributable to the treatment of a large enough magnitude to be considered clinically significant? and, Are the treated individuals distinguishable from normal individuals with regard to target variables? It is important to note that these two questions are somewhat independent. The first question addresses the magnitude of the change attributed to treatment, whereas the second addresses the impact of that change. Both issues are important to the issue of clinical significance (Kazdin & Weisz, 1998). For example, a successful treatment of individuals with severe psychopathology may lead to change of large magnitude while failing to return treated individuals to within normal limits. Few would argue that a treatment that is able to effect an 80% reduction in psychotic symptoms among schizophrenics is unsuccessful, even though the remaining symptomatology at post-treatment may still fall outside the normal range. Thus, for individuals with certain types of severe psychopathology, normative comparisons offer a stringent criterion. Conversely, a treatment that returns treated individuals to within normal limits but does not exhibit clinically significant change may be viewed with suspicion. This scenario begs the question, How severe were the client's symptoms to begin with?

Separate methodologies may be brought to bear on each of these questions. Specifically, Jacobson, Follette, and Revenstorf (1984) developed a technique known as the reliable change index that may be used to determine whether a treated individual has undergone clinically significant, reliable change. More recently, this technique has been updated (Jacobson & Truax, 1991) and adapted for use on samples rather than individuals so that the RCI may be now used to evaluate entire treatment studies (Abramowitz, 1998).

Likewise, Kendall and Grove (1988) described a technique known as normative comparisons for use in therapy outcome research. Recently, the statistical technique of equivalency testing (Rogers, Howard, & Vessey, 1993) has been applied to the task of normative comparisons, making statistical interpretation possible (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999) and shedding light on the question of whether the functioning of a participant group after treatment should be regarded as having been returned to normative levels. For the purposes of this review, this approach rests on the idea that symptomatology related to conduct problems in children varies in the general population, with most children exhibiting at least some level of problem behavior. If post-treatment group means can be reliably brought either close to or below the normative mean, then evidence has been gained that

supports the clinical significance of the post-treatment results.

METHODS

Studies and Measures

Use of normative comparisons and the RCI to assess clinical significance both require a common outcome variable. Ideally, the outcome variable should be widely used and possess ample normative data. Of the 82 studies identified by Brestan and Eyberg (1998), 30 studies evaluated 1 of the 12 different treatments identified as possessing empirical support. Among these 30 studies, the most commonly used outcome measures were the Child Behavior Checklist (CBCL; Achenbach & Edelbrook, 1983) and the Eyberg Child Behavior Inventory (ECBI; Eyberg & Ross, 1978). In addition to being widely used, both of these measures are well validated and possess a large amount of normative data sensitive to age level (Achenbach, 1991; Burns & Patterson, 1990; Burns, Patterson, Nussbaum, & Parker, 1991). Furthermore, retest reliability is high for both the CBCL (r = .952; Achenbach, 1991) and the ECBI (r = .86; Eyberg & Ross, 1978). For these reasons, the CBCL and the ECBI, specifically the intensity and problems subscales on the ECBI (ECBI-I and ECBI-P) and the total problem and externalizing subscales on the CBCL (CBCL-TP and CBCL-E), were chosen as the measures for examination in this study.

Of the 30 studies cited by Brestan and Eyberg (1998), 10 studies were identified that used either the CBCL or the ECBI. One study (Spaccarelli, Cotler, & Penman, 1992) was excluded due to an insufficient level of reported data. The remaining nine studies evaluated three treatment approaches. On this basis, the three treatments chosen for review were Webster-Stratton's Videotape Modeling (VM; Webster-Stratton, 1984) Parent Training, Kazdin's Problem-Solving Skills Training (PSST; Kazdin, Bass, Siegal, & Thomas, 1989), and Eyberg's Parent-Child Interaction Therapy (PCIT; Eyberg, Boggs, & Algina, 1995). In addition, five further studies were identified as having used at least one of these measures. These five studies were not included in the Brestan and Eyberg (1998) review because of the recency of their publication. Demographic information and inclusion criteria for each study are provided in Table 1.

For each outcome study, treatment groups were included based on inclusion of the primary techniques studied by the investigator; that is, VM, PSST, and PCIT. Alternative treatment groups were excluded if one of these three techniques was not part of the treatment provided. For example, Taylor, Schmidt, Pepler, and Hodgins (1998) compared parenting groups to a routine-care condition consisting of an eclectic array of approaches found at a local treatment center. In this case, the routine-care condition was not included in the present review.

Because inclusion in this review was based on the presence of one of the primary treatment techniques, each treatment modality reviewed in this paper covers a range of implementations. For example, VM includes group and individual formats as well as treatments that are directed at parents and children independently and in combination. PSST has been studied both alone and in combination with parent management training. Differing implementations of a particular approach were not compared due to lack of power.

In summary, 14 studies of 3 treatments are reviewed. Some of these studies compared two different implementations of the same approach. Thus, among these 14 studies, 23 treatment groups are evaluated. Furthermore, the progress of each treatment group was evaluated by between one and three of the measures considered herein (ECBI-I, ECBI-P, CBCL-E, and CBCL-TP). Thus, 1 of these 4 measures was used in a total of 50 instances (see Table 2).

For the purposes of this review, when data were provided by multiple informants (i.e., mother, father, and teacher), only data from the mother are included. This decision was made because mother-reported data were by far the most commonly cited among the 15 studies considered. In addition, those studies that did include paternal-report data typically reported smaller sample sizes among fathers than among mothers. Thus, choosing maternal-report maximized power. Furthermore, compared to the average level of inter-parent agreement reported in a meta-analysis of cross-informant agreement (r = .60; Achenbach, 1997), inter-parent agreement is average to high on both the ECBI (r = .61 on the problem scale and r = .69 on the intensity scale; Eisenstad, McElreath, Eyberg, McNeil, & Newcomb, 1994) and on the CBCL (r = .76 on the total problems scale and r = .81 on the externalizing scale; Achenbach, 1991). Although parents agree more about their children's conduct problems then about most child attributes, they still tend to agree only moderately about their children's conduct. Thus, even though restricting the review to maternal

Table 1. Demographics and inclusion criteria for selected studies

Study	*N*	Age range, Years (*M*)	% Male	Ethnicity	Inclusion Criteria (Sample Characteristics)
Taylor et al. (1998)	16	3–8 (5.6)	73.9	Primarily Northern European descent	Referral to community-based treatment center (85% exceeded clinical cutoff for ECBI problem scale)
Webster-Stratton (1984)	31	3–8 (4.7)	71.4	NR	Referral for oppositional behavior
Webster-Stratton et al. (1988)	54	3–8 (4.5)	69.3	NR	Primary referral of child misconduct; exceed clinical cutoff on ECBI problem scale
Webster-Stratton (1990)	33	3–8 (5.1)	79.1	NR	Primary referral of child misconduct; exceed clinical cutoff on ECBI problem scale (mean ECBI problem = 20.6)
Webster-Stratton (1994)	77	3–8 (4.9)	74.4	NR	Primary referral of child misconduct; diagnosis of ODD and/or CD; exceed clinical cutoff on ECBI problem scale
Webster-Stratton & Hammond (1997)	75	4–7 (5.7)	74.2	86% Caucasian	Primary referral of child misconduct; diagnosis of ODD and/or CD; exceed clinical cutoff on ECBI problem scale
Kazdin et al. (1987a)	24	7–12 (10)	77.5	75% Caucasian, 25% African American	Referral for antisocial behavior; inpatient status (2–3 months); ≥ 98th %tile on CBCL aggression or delinquency scales (78% diagnosed CD)
Kazdin et al. (1987b)	20	7–13 (10.9)	79	77% Caucasian, 23% African American	Referral for antisocial behavior; inpatient status (2–3 months); ≥ 98th %tile on CBCL aggression or delinquency scales
Kazdin et al. (1989)	75	7–13 (11)	78	55% Caucasian, 46% African American	≥ 90th %tile on CBCL aggression or delinquency scales (75% diagnosed CD)
Kazdin et al. (1992)	66	7–13	78	69% Caucasian, 31% African American	≥ 90th %tile on CBCL aggression or delinquency scales (50% given primary diagnosis of CD)
Eisenstadt et al. (1993)	24	2.5–7	92	88% Caucasian	Diagnosis of ODD, CD, or ADHD (88% were diagnosed with ODD or CD)
Eyberg et al. (1995)	10	(4.5)	80	80% Caucasian	Diagnosis of ODD (33% were comorbid ADHD)
McNeil et al. (1991)	10	2–7 (4.5)	100	90% Caucasian	ECBI Intensity and Problem scores above published cutoff scores for deviancy; RCTRS score > 1 standard deviation above mean (100% ODD, 90% ADHD; 30% CD)
Schuhmann et al. (1998)	28	3–6 (4.9)	81	77% Caucasian, 14% African American	ODD diagnosis (66% ADHD, 22% CD)

Notes. CD = conduct disorder; ODD = oppositional defiant disorder; ADHD = attention-deficit hyperactivity disorder; CBCL = Child Behavior Checklist; ECBI = Eyberg Child Behavior Inventory; RCTRS = Revised Conners Teacher Rating Scale; NR = not reported.

reports is not likely to misrepresent the data, some richness will invariably be lost.

Treatment Methods

The VM parent-training program was identified by Brestan and Eyberg (1998) as a well-established treatment. It includes a cost-effective videotape series of parent-training lessons and is based on Hanf's (1969) model of parent training. Like other forms of parent management training, the goal of VM is to teach parents specific procedures to alter interactions with their child, promote prosocial behavior, and decrease deviant behavior, particularly in the home (Kazdin & Weisz, 1998). Therapists use a collaborative approach to foster discussion of the videotape vignettes to accomplish these goals (Webster-Stratton & Herbert, 1993). The treatment is designed for young children (generally ages 3–8 years) with a range of problems, from mild noncompliance to full conduct disorder.

The second treatment, PSST, was identified by Brestan and Eyberg (1998) as a probably efficacious treatment. Several elements are common to all implementations of this approach. It is targeted at school-age children, and emphasis is placed on children's cognitive approach to situations, particularly interpersonal problems. Prosocial responses to these situations are modeled and reinforced through games, academic activities, stories, and increasingly as treatment progresses, real-life situations. Thera-

Table 2. Characteristics, pre- and post-treatment scores, RCI, and normative comparisons for each treatment group

Study	Treatment	*N*	Measure	Pretest *M* (*SD*)	Post-test *M* (*SD*)	RCI[a]	Normative Comparison[b]
Taylor et al. (1998)	Parenting group VM	16	ECBI-P	19.3 (19.3)	12.4 (7.1)	1.57	Different
			ECBI-I	139.3 (36.4)	115.2 (27.3)	1.25	Equivalent
			CBCL-TP	54.4 (35)	37.7 (20.5)	1.54	Different
Webster-Stratton (1984)	Group therapy VM	15	CBCL-TP	61.7 (18.2)	37 (14.2)	4.37*	Different
			ECBI-P	19.5 (7.9)	8.7 (6.2)	2.58*	Equivalent
			ECBI-I	144 (30.3)	102.9 (22.2)	2.57*	Equivalent
	Parent & child VM	16	CBCL-TP	71 (33.9)	42.7 (29.7)	2.7*	Different
			ECBI-P	22.4 (5.6)	9.1 (8.6)	4.47*	Equivalent
			ECBI-I	166.6 (22.9)	115.1 (19.3)	4.25*	Equivalent
Webster-Stratton et al. (1988)	Individual VM	27	ECBI-P	20.1 (5.8)	11.7 (5.7)	2.74*	Different
			ECBI-I	156.1 (23.9)	126.2 (23.2)	2.37*	Different
			CBCL-TP	61.1 (26.4)	38 (21.6)	2.83*	Different
	Group therapy VM	27	ECBI-P	21 (5.4)	12.8 (8.4)	2.88*	Different
			ECBI-I	159 (25.4)	111.1 (33.4)	3.56*	Equivalent
			CBCL-TP	53.2 (14.2)	31.1 (18.8)	5.03*	Equivalent
Webster-Stratton (1990)	Parent VM	17	CBCL-TP	49.3 (19.2)	37.6 (18.4)	1.97*	Different
			ECBI-I	164.6 (29)	123 (28.3)	2.7*	Different
	Parent & child VM	16	CBCL-TP	64.5 (21.6)	45.2 (22.7)	2.89*	Different
			ECBI-I	155.5 (17.1)	129.1 (26.2)	2.92*	Different
Webster-Stratton (1994)	Group therapy VM	39	CBCL-TP	64.1 (8.6)	57.8 (9.6)	2.37*	Different
			ECBI-P	21.3 (5.7)	12.5 (6.5)	2.94*	Different
	Group therapy VM+	38	CBCL-TP	66.2 (9)	57.5 (11.1)	3.14*	Different
			ECBI-P	21.2 (5.3)	8.7 (6.4)	4.4*	Equivalent
Webster-Stratton & Hammond (1997)	Parent VM	26	CBCL-TP	65.5 (7.8)	56 (8.9)	3.93*	Equivalent
			ECBI-I	166.5 (23.7)	118.7 (27.7)	3.8*	Equivalent
	Child VM	27	CBCL-TP	67.1 (8)	62.2 (9.4)	2*	Different
			ECBI-I	155.5 (29.1)	121.7 (23)	2.2*	Different
	Parent & child VM	22	CBCL-TP	65.3 (6.1)	57.1 (7.7)	4.36*	Different
			ECBI-I	161.6 (33.4)	121.4 (24.3)	2.27*	Different
Kazdin et al. (1987a)	PSST + parent management	24	CBCL-E	77.6 (5.8)	66.7 (6.1)	6.07*	Different
			CBCL-TP	78.9 (8.9)	66.4 (7.4)	4.53*	Different
Kazdin et al. (1987b)	PSST	20	CBCL-E	78.5 (4.9)	66.8 (9.4)	7.71*	Different
			CBCL-TP	79 (5.5)	67.5 (9.8)	6.75*	Different
Kazdin et al. (1989)	PSST	37	CBCL-E	77.8 (7.6)	67.9 (8.9)	4.2*	Different
			CBCL-TP	77.5 (8.4)	67 (10.6)	4.03*	Different
	PSST + in vivo practice	38	CBCL-E	79 (6.9)	65 (10.7)	6.55*	Different
			CBCL-TP	78.2 (9)	64.2 (10.8)	5.02*	Different
Kazdin et al. (1992)	PSST	29	CBCL-TP	72 (8.4)	64.6 (8.5)	2.84*	Different
	PSST + parent management	37	CBCL-TP	69.6 (7.5)	60.2 (10.7)	4.05*	Different
Eisenstadt et al. (1993)	PCIT	24	ECBI-I	173 (29.1)	101.8 (23.3)	4.62*	Equivalent
			ECBI-P	23 (6)	5.6 (6)	5.48*	Equivalent
			CBCL-E	73.4 (7.5)	61.2 (7.6)	5.25*	Different
Eyberg et al. (1995)	PCIT	10	ECBI-P	20.7 (4.8)	6.6 (6.7)	5.55*	Equivalent
			ECBI-I	159.5 (16.6)	117.5 (18.8)	4.78*	Equivocal
McNeil et al. (1991)	PCIT	10	ECBI-P	23.3 (6.7)	6.1 (7.7)	4.85*	Equivalent
			ECBI-I	180.7(28.2)	105.9 (29.2)	5.01*	Equivalent
Schuhmann et al. (1998)	PCIT	22	ECBI-P	21.9 (6.5)	10.9 (9.6)	3.2*	Different
			ECBI-I	170.3 (26.4)	117.6 (40.4)	3.77*	Equivalent
	PCIT	6	ECBI-P	21.3 (6)	9.8 (10.7)	3.62*	Equivocal
			ECBI-I	160.3 (22.5)	127.3 (22.9)	2.77*	Different

Notes. VM = videotape modeling; PSST = Problem-Solving Skills Training; PCIT = Parent-Child Interaction Therapy; ECBI = Eyberg Child Behavior Inventory (I = intensity; P = problems subscales); CBCL = Child Behavior Checklist (TP = total problem; E = externalizing subscales).
[a]Significant RCI values (RCI > 1.96) are marked with an asterisk.
[b]Equivalent = significant equivalency test; different = significant traditional *t*-test and nonsignificant equivalency test; equivocal = non-significant equivalency test and nonsignificant traditional *t*-test.

pists assume an active stance, often modeling the problem-solving skills presented through self-statements (Kazdin & Weisz, 1998). PSST is unique among treatments in this review in that it targets adolescents, whereas the other two treatments are directed at younger children.

The third treatment, PCIT, was also identified by Brestan and Eyberg (1998) as probably efficacious and has been reviewed most prominently by Eyberg et al. (1995). Like VM, PCIT is also based on Hanf's (1969) model of parent-training and directed at young children. It thus

shares the emphasis on teaching parents specific procedures to alter interactions with their child. However, it differs from VM in its emphasis on conducting treatment through instruction during parent-child interactions (McNeil, Eyberg, Eisenstadt, Newcomb, & Funderburk, 1991), rather than instructing parents or children during individual meetings with a therapist.

RESULTS

Data were analyzed using RCI and normative comparisons. In each case, we first compared measures to determine whether there is any evidence that a finding of clinical significance is more likely on either the ECBI or on the CBCL. Second, the three treatments were evaluated and compared on their ability to attain clinically significant results.

Clinically Significant Change

To determine whether the change effected by a given treatment method was of a large enough magnitude to be considered clinically significant, the RCI, as adapted for group data by Abramowitz (1998), was used.[1]

Comparison of Measures. Of the 23 treatment conditions reviewed, 15 were evaluated with both the ECBI and the CBCL. Among these groups, 13 demonstrated clinically significant change on both measures, and two exhibited clinically significant change on neither measure. Thus, a McNemar test of dependent proportions was not necessary. There was not a single instance of a treatment attaining clinically significant change on one measure but not the other, and thus the discordant cells of a McNemar test would be equal to zero. Therefore, there is no evidence that it is easier to obtain a result of clinically significant change using any scale of the ECBI or the CBCL based on the studies reviewed in this article.

Evaluation of Treatments. Overall, the three treatments reviewed were highly successful at eliciting clinically significant change as indexed by the RCI (see Table 2). Of the 50 instances in which the ECBI or the CBCL was used in this review, 46 (92%) were associated with clinically significant change. Twenty-two of the 23 treatment conditions (96%) reviewed had at least 1 post-treatment mean that demonstrated clinically significant change.

Treatments were also compared. Of the 29 evaluations of VM, 25 (86%) were associated with clinically significant change. All 10 evaluations of PSST yielded a finding of clinically significant change. Likewise, all 11 evaluations of PCIT yielded clinically significant results. A multiple independent-samples chi-square test of these proportions was not significant ($\chi^2_2 = 0.13$, $N = 50$, $p > .1$). This result indicates that there is insufficient evidence to support the hypothesis that differences exist among the three treatments in regard to their ability to effect clinically significant change as indexed by the RCI.

With the CBCL-TP and CBCL-E as the sole basis of comparison, Webster Stratton's VM yielded 12 out of 13 (92%) clinically significant results, PSST yielded 6 out of 10 (60%) clinically significant results, and PCIT yielded a clinically significant result on the one occasion that the CBCL-E was used. A multiple independent samples chi-square test of these three proportions was not significant ($\chi^2_2 = 0.45$, $N = 24$, $p > .1$), indicating that there were no detectable differences among the treatments based on the CBCL alone. Using the ECBI-I and ECBI-P as the basis of comparison, Webster Stratton's VM yielded 16 out of 19 (84%) clinically significant results, and PCIT yielded a clinically significant result on the all 10 occasions that the ECBI-I or ECBI-P was used. A chi-square test of these two proportions was not significant ($\chi^2_2 = 0.09$, $N = 29$, $p > .25$), indicating that there was no significant difference between Webster Stratton's VM and PCIT based on the ECBI alone. These data lend further support to the conclusion that there are no significant differences among the three treatments in regard to their ability to produce clinically significant change on the RCI.

Normative Comparisons

To determine whether treated individuals were returned to within normal limits on a given outcome variable, equivalency testing as described by Kendall, Marrs-Garcia, et al. (1999) was used. First, a range of closeness around the normative mean within which a subject group's mean would be deemed clinically equivalent was specified. For the purposes of this study, one standard deviation around the normative mean was used. This value was chosen to correspond to the average range of the CBCL (Achenbach, 1991) and the normative range of the ECBI (Eyberg & Ross, 1978) as reported in validation studies of these measures using normative samples.

Next, equivalency tests and traditional tests of means were conducted to evaluate normative comparisons.[2] Results were classified as either "equivalent" (correspond-

Table 3. Classification of results from tests of the statistical significance of effects and clinical equivalency tests

Clinical Equivalency Test	Traditional Statistical Tests: Significant Effect	Traditional Statistical Tests: Nonsignificant Effect
Significant	*Cell I:* Statistically different; clinically equivalent	*Cell II:* Clinically equivalent
Not significant	*Cell III:* Different (not clinically equivalent)	*Cell IV:* Equivocal findings (more power required)

ing with cells I and II; see Table 3), "different" (corresponding to cell III), or "equivocal" (corresponding to cell IV). "Equivalent" indicates that the post-treatment means fall reliably within one standard deviation of the normative mean and are thus considered to be clinically equivalent to the norm. Results falling in the cell II, "clinically equivalent," or cell I, "statistically different, clinically equivalent" are indicative of clinically significant results. "Different" indicates that the post-treatment results are statistically different from the norm. In the absence of a finding of equivalency (i.e., cell III), a finding of statistical difference indicates that the treatment mean falls reliably outside of one standard deviation of the normative mean. "Equivocal" indicates that findings are not interpretable as either "different" or "equivalent," perhaps due to lack of power.

Normative Samples. Studies of normative samples of the CBCL and the ECBI show significant effects due to age (Achenbach, 1991; Burns & Patterson, 1990; Burns et al., 1991). For this reason, normative data for the CBCL and the ECBI were chosen so as to most closely approximate the ages of each particular subject group. For the CBCL, only two broad age ranges are offered, and studies typically report data in the form of T-scores, so this objective presented little difficulty (when raw data were reported, raw data from normative samples were used for comparison). However, the two sources of normative samples considered for the ECBI provided data for much smaller age ranges that seldom corresponded precisely to the age ranges of the subject groups included in the treatment studies. When a subject group's age range covered more than one normative sample, normative groups for relevant age ranges were combined to form an age-appropriate normative comparison group. A weighted mean and a pooled variance was calculated for each group.[3]

Comparison of Measures. Is there any evidence that there is a greater probability of obtaining a finding of clinical significance based on the ECBI than on the CBCL? To answer this question, we analyzed treatments that were evaluated using both the ECBI and the CBCL. Of the 12 treatment groups evaluated with both the ECBI-I and the CBCL-TP, 2 demonstrated clinical equivalency to the norm at post-treatment on both measures, 5 on neither measure, 1 on only the CBCL, and 4 on only the ECBI-I. To test the hypothesis that the proportion of clinically equivalent results was the same for both the ECBI-I and the CBCL-TP, a McNemar Z_{UN} test of dependent proportions was performed. The result was not significant ($Z_{UN} = .89$, $p > .18$). Of the 10 treatment groups evaluated with both the ECBI-P and the CBCL-TP, 1 demonstrated equivalency to the norm at posttreatment on both measures, 3 on neither measure, 1 on only the CBCL, and 5 on only the ECBI-P. To test the hypothesis that the proportion of clinically equivalent results was the same for both the ECBI-P and the CBCL-TP, a McNemar Z_{UN} test of dependent proportions was performed. Again, the result was not significant ($Z_{UN} = 1.22$, $p > .10$), indicating that there is not enough evidence to reject the hypothesis that obtaining a finding of clinical equivalency is easier on one measure than on the other. However, a post-hoc power analyses indicated that this result would have been significant with 20 observations rather than 10.[4] For these studies, the probability of obtaining a result of clinical equivalency on the ECBI given a non-significant result on the CBCL-TP was $Pr = .5$, while the probability of obtaining a finding of clinical equivalency on the CBCL-TP given a nonsignificant result on the ECBI was $Pr = .14$. Thus, the nonsignificant McNemar test should be viewed with suspicion as an underpowered test, and results that demonstrate a treatment's inability to obtain findings of clinical equivalency based exclusively on the CBCL-TP should be interpreted with caution.

Evaluation of Treatments. Overall, when using normative comparisons, the three treatments reviewed were moderately successful at returning groups to within normal limits on the ECBI and CBCL (see Table 2). Of the 50 instances in which the ECBI or the CBCL was used, 20 (40%) indicated that the post-treatment mean was clinically equivalent to the norm. Eleven of the 23 treatments (48%) had at least 1 post-treatment mean that was clinically equivalent to the norm.

Treatments were also compared using the normative

comparison approach to clinical significance. Out of 29 evaluations of VM, the post-treatment mean was clinically equivalent to the norm for 13 (45%). Seven out of 11 evaluations (64%) of PCIT yielded post-treatment means that were clinically equivalent to the norm. Finally, of 10 evaluations of PSST, none brought the post-treatment mean to within normal limits. A multiple independent samples chi-square test of these proportions was significant ($\chi^2_2 = 5.06$, $N = 50$, $p<.05$). This result indicates that differences exist among the proportions for the three treatments. Post-hoc chi-square tests with Bonferonni corrections for the number of tests were conducted for each pair of proportions. The results comparing VM and PCIT were not significant, ($\chi^2_1 = .31$, $N = 40$, $p>.05$). However, the results comparing PSST with PCIT ($\chi^2_1 = 5.19$, $N = 21$, $p<.0167$) and with VM ($\chi^2_1 = 4.13$, $N = 39$, $p<.0167$) were both significant. Thus, although no significant differences were found between PCIT and VM, both of these treatments were more likely than PSST to produce post-treatment means that were clinically equivalent to the norm (see Discussion for a client characteristic that may explain this finding).

Using the ECBI-I and ECBI-P as the basis of comparison, VM yielded 10 out of 19 (53%) post-treatment means that were clinically equivalent to the norm, and PCIT yielded post-treatment means that were clinically equivalent to the norm on 7 of 10 occasions (70%). The chi-square test of these two proportions was not significant ($\chi^2_1 = 0.21$, $N = 29$, $p>.5$).

For PSST, a lower proportion of post-treatment means were clinically equivalent to the norm at post-treatment. Given this result, the question arose of whether there is any evidence that those treatment groups whose means were not equivalent to the norm at post-treatment were composed of participants with more severe initial conduct problems, as reflected by higher pretreatment scores. To answer this question, pretreatment means of those groups that demonstrated clinical equivalency to the norm were compared to the pretreatment means of those groups that did not demonstrate clinical equivalency. For the ECBI-P, the grand mean of pretreatment scores for those studies that obtained clinical equivalency was actually higher than the grand mean of the pretreatment scores for those studies that did not obtain clinical equivalency ($M = 21.58$ and $M = 20.39$, respectively). The same held for the ECBI-I ($M = 161.52$ and $M = 158.02$, respectively). For the CBCL-TP, the pretreatment means of clinically equivalent results ($M = 66.5$, $n = 3$) were compared to the pretreatment means of results that were not clinically equivalent ($M = 71.1$, $n = 17$) using a *t*-test for independent samples. The result was not significant ($t_{18} = .1$, $p > .5$), and the effect size was very small (.06). Thus, there is little evidence to conclude that those groups whose means were not equivalent to the norm at post-treatment possessed higher pretreatment means than those groups that did obtain clinical equivalency.

DISCUSSION

This article outlines one approach to the analysis of clinical significance and demonstrates that two approaches to clinical significance, the RCI and normative comparisons, offer information that complements one another as well as the results of RCTs. Although past RCTs and meta-analyses have demonstrated the efficacy of the three treatments for conduct-disordered children we reviewed (VM, PSST, and PCIT), the clinical significance of improvements attributed to these treatments has not been previously explored. The present results indicate that the clinical significance of the results of the three treatments reviewed was high. Approximately 96% of all treatment conditions demonstrated clinically significant change on at least one measure of child behavior, and approximately 48% of treatment groups were clinically equivalent to the norm on at least one measure of child behavior at post-treatment.

Direct comparisons were made among the three treatments reviewed. No significant differences were detected among the treatments' abilities to produce clinically significant change using the RCI. The equivalence of such diverse treatments targeted at different populations may at first seem curious. However, studies were included specifically because they investigated treatments that have already been found to be effective. Given that all three treatments consistently produce statistically significant change, it is not completely surprising that in general they also produce clinically significant change on the same measures. In this regard, the present findings generally confirm what is to be expected of three well-accepted treatments.

However, the exceptions are instructive. In terms of normative comparisons, both PCIT and VM were more likely than PSST to produce post-treatment means that were clinically equivalent to the normative population. These results must be interpreted with caution, however, due to the likelihood of a lesser probability of obtaining clinical equivalency using the CBCL, which was used

exclusively in studies of PSST. Perhaps more important, assuming differences among the treatments do exist, the results could be attributed either to the treatment or to the treated populations. PSST differs significantly from the other two treatments and is directed at a significantly older population of children (mean age = 10.8 years) than either VM (mean age = 5.2 years) or PCIT (mean age = 4.6 years). Moreover, several studies of PSST were conducted with inpatient populations, probably indicating a much more severe level of pathology. Thus, the observed differences among treatments may be a reflection of the nature of conduct problems in children; as children with conduct problems grow older, it may become increasingly more difficult to return their behavior to normative levels, although clinically significant change may remain possible. This interpretation of the data is consistent with other reviews of the literature that have found child age to be a significant predictor of attrition in studies of parent training (Dishion & Patterson, 1992; Southam-Gerow & Kendall, 1997).

As stated above, there may be differences between the ECBI and the CBCL in their ability to produce significant results with normative comparisons. Why the difference? Although the two measures are both behavior inventories, the ECBI includes a severity index for each item. Thus, whereas participants only have a range of three choices for each item on the CBCL, on the ECBI they may indicate that they observe a certain behavior and then rate its interference on a scale from one to seven. The difference between the two measures in terms of sensitivity to the severity of specific behavior problems may explain the present results. In any case, the difference between two such similar measures in terms of normative comparisons highlights the importance of paying careful attention to measurement when considering clinical significance.

A potential limitation of this review is that the selection of studies was based on their use of two specific outcome measures. As a result, one treatment method identified by Brestan and Eyberg (1998) as "well-established" and seven identified as "probably efficacious" were not included in this study. These included treatments such as anger coping therapy (e.g., Lochman, Burch, Curry, & Lampron, 1984), multisystemic therapy (e.g., Borduin et al., 1995), and those based on Patterson's Living with Children (Patterson, 1976). Furthermore, due to lack of uniform use of specific measures and subscales even within the present sample of studies, direct comparisons among treatment methods were difficult. Consistency in the use of outcome measures would be beneficial to better conduct cumulative reviews of the outcome literature.

Related limitations derive from reliance on a subset of outcome measures. Both measures used were based solely on maternal report. Additionally, the CBCL represents a behavior checklist that does not consider the impact or interference caused by specific problem behaviors. It is possible that, although problem behaviors tended to be greatly reduced, interference remained high, especially in contexts outside of the home. Alternatively, behavior checklists may not adequately represent the results of treatments that target a subset of problem behaviors responsible for the highest levels of interference and disruption. For example, a treatment that successfully reduces the number of fights a conduct-disordered boy starts and also brings that boy back to school may drastically improve quality of life for the family and be considered a success. However, if other problem behaviors persist, such gains will not lead to much change in behavior checklist scores. Treatments that fell short of obtaining clinical significance on the ECBI or the CBCL thus may have met with more success on different measures or on those derived from different informants.

Finally, it should be noted that the methods used in this analysis represent only a subset of the possible approaches to clinical significance. For example, although group data may be most appropriate for cumulative reviews of treatment studies (given the nature of published data), individual data regarding clinical significance may also be appropriate for randomized clinical trials. Furthermore, we focused on measures of symptomatology. Clearly, in terms of clinical significance, overt symptoms are only part of the picture. As many investigators of antisocial behavior have pointed out (e.g., Hengeller et al., 1986; Hengeller, Melton, & Smith, 1992), other important variables, such as quality of life, school performance, arrest and incarceration records, and measures of the quality of familial and peer relationships, may be of equal or greater interest. For each of these variables, similar questions arise. Was the change in school performance clinically significant? Were the peer relationships of treated individuals similar to nondisordered individuals of the same age? Such questions can be explored with the method proposed here.

Overall, results of the method proposed here should be interpreted within the context of the measures employed.

For example, we considered ECBIs and CBCLs filled out by mothers of referred children. "Clinically significant change" in this case meant that mothers saw significant improvement in the problem behavior of their children. "Clinical equivalence" means that these mothers reported about as many problem behaviors post-treatment as did mothers of normal children. These results say nothing about other contexts. They do not, for example, mean that the target children necessarily did better in school or had more friends or better peer relationships. A wider range of measures would need to be analyzed with the proposed method to address such questions.

The present review encourages use of the RCI (as adapted for groups) and the normative comparisons methodology for cumulative evaluations of treatment studies. For the studies reviewed in this article, each method provided unique information regarding clinical significance, lending support to the idea that these two methods may most profitably be used in conjunction. Together, the methods offer information to benefit researchers and clinicians alike. In the research vein, they may be used to help establish empirical support for treatments, thus adding a valuable dimension to the criteria discussed by Chambless and Hollon (1998). From a clinical perspective, use of the methods may help a clinician familiar with the research literature determine what they can in good faith tell a prospective client about the track record of a treatment method that has already obtained empirical support. Is the change expected to result from treatment likely to be meaningful? Will clients be returned to within a normal range after treatment with respect to their primary complaints? With the aid of the RCI and normative comparisons, information can be gleaned to help the clinician formulate answers, and the answers may be somewhat specific. It is not unlikely that on some variables, individuals will be within the normal range, such as on skills directly targeted by treatment. On other variables, mean post-treatment levels may not achieve of normative levels. In this case, clinically significant change would indicate that although the treatment is by no means a failure, the expectations of the clinician may need to be tempered. For example, it appears from this review that with respect to child behavior problems, obtaining equivalency with normative populations may represent the more rigorous criterion, particularly with respect to older groups of children. However, the consistent findings of reliable, clinically significant change indicate that treatment is worthwhile, even if equivalency with normative levels of behavior cannot always be attained.

NOTES

1. The RCI is calculated for each outcome measure of each treatment condition as follows:

$$\mathrm{RCI} = \frac{M_{\mathrm{pre}} - M_{\mathrm{post}}}{S_{\mathrm{diff}}}$$

where M_{pre} and M_{post} are the mean pre- and post-treatment scores for the treatment condition on one of the ECBI or CBCL subscales. S_{diff} is the standard error of the difference between the two means and describes the distribution of change scores that would be expected if actual change did occur (Jacobson & Truax, 1991). S_{diff} is calculated as:

$$S_{\mathrm{diff}} = \sqrt{2(SE)^2};\ SE = SD_{\mathrm{pre}}\sqrt{1 - r_{xx}}$$

where SE is the standard error of measurement, SD_{pre} is the standard deviation of the pretreatment score, and r_{xx} is the retest reliability of the outcome measure. At $p<.05$, an RCI of 1.96 or greater indicates that the change attributable to treatment on a given outcome measure is most likely not due to chance and may be considered reliable and clinically significant.

2. For each treatment condition, two one-tailed z-tests were conducted to determine whether the mean and the normative mean were sufficiently close to be considered clinically equivalent. Because the specified range of closeness in this case is symmetrical about the normative mean, only one of these tests was conducted:

$$z = \frac{M_{\mathrm{N}} - M_{\mathrm{C}} - \delta}{\sqrt{\left(\frac{(n_{\mathrm{N}} - 1)SD_{\mathrm{N}}^2 + (n_{\mathrm{C}} - 1)SD_{\mathrm{C}}^2}{n_{\mathrm{N}} + n_{\mathrm{C}} - 2}\right)\left(\frac{1}{n_{\mathrm{N}}} + \frac{1}{n_{\mathrm{C}}}\right)}}$$

where M_{N} is the normative mean, M_{C} is the clinical mean, δ is the specified one standard-deviation range of closeness, n_{N} is the normative sample size, n_{C} is the clinical sample size, SD_{N} is the standard deviation of the normative sample, and SD_{C} is the standard deviation of the clinical sample. At $p<.05$, a one-tailed z-score of 1.64 or greater indicates that the population mean corresponding to the subject group most likely lies within the specified range of closeness to the normative mean.

Z-tests for equivalence were conducted in conjunction with traditional z-tests for differences between means. A significant result on the latter test indicated that the difference between the two means was statistically significant. The information from both tests was combined as illustrated in Table 3 (Kendall, Marrs-Garcia, et al., 1999). As can be seen in the Table 3, significant results on only the equivalency test lead to a finding of "clinical equivalence," indicating that the post-treatment mean is equivalent to the normative group. Significant results on only the tradi-

tional test leads to a finding of "difference," indicating that the posttreatment mean is statistically different from the normative group. Significant results on both tests lead to a finding of "statistical difference and clinical equivalency." Because these results reliably fall within the normative range, they are classified in Table 2 as equivalent. Significant results on neither test leads to "equivocal findings," suggesting that more power is required to obtain interpretable results.

To rule out instances where treatment means were equivalent to the norm even at pretreatment, both equivalency testing and traditional hypothesis testing was conducted for both pre- and post-treatment means. Only those measures that moved from the category "different" to either of the "clinically equivalent" categories were considered to be clinically equivalent to the norm for the purposes of this study. However, moving from one category to another does not suggest that the change in scores from pre- to post-treatment is significant; this criterion simply excludes cases that remain equivalent to normative levels at both pre- and post-treatment. All treatment groups considered in this analysis were different at pretreatment on the measures reported.

3. Equations used were as follows:

$$M_N = \frac{n_1 M_1 + n_2 M_2 + \ldots}{n_1 + n_2 + \ldots};$$

$$\sigma = \sqrt{\frac{(n_1 - 1)SD_2^2 + (n_2 - 1)SD_2^2 + \ldots}{(n_1 - 1) + (n_2 - 1) + \ldots}}$$

where M_N is the combined normative mean, n_x is the size of one of the normative samples, M_x is the mean of the sample, SD_x is the standard deviation of the sample, and σ is the pooled standard deviation of all normative samples. M_N and σ were used in all calculation involving normative samples.

4. To determine n needed to obtain a statistically equivalent result, the McNemar equation, $Z_{un} = 1.96 = \frac{(5x - x) - 1}{\sqrt{5x + x}}$ was solved using the quadratic equation, yielding a result of $x = 1.91$. Multiplying the 10 observations included in the McNemar test by 1.91, we obtain a result of 19.1 observations, indicating that a sample size of at least 20 would have been required to obtain a significant result.

REFERENCES

Abramowitz, J. S. (1998). Does cognitive-behavioral therapy cure obsessive-compulsive disorder? A meta-analytic evaluation of clinical significance. *Behavior Therapy, 29,* 339–355.

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4–18 and 1991 Profile.* Burlington, VT: University of Vermont Department of Psychiatry.

Achenbach, T. M. (1997). What is normal? What is abnormal? Developmental perspectives on behavioral and emotional problems. In S. Luthar, J. Burack, D. Cicchetti, & J. Weisz (Eds.), *Developmental Psychopathology: Perspectives on adjustment, risk and disorder* (pp. 93–114). New York: Cambridge University Press.

Achenbach, T. M., & Edelbrook, C. S. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile.* Burlington, VT: University Associates in Psychiatry.

Bank, L., Duncan, T., Patterson, G. R., & Reid, J. (1994). Parent and teacher ratings in the assessment and prediction of antisocial and delinquent behaviors. *Journal of Personality, 61,* 693–709.

Borduin, C. M., Mann, B. J., Cone, L. T., Henggeler, S. W., Fucci, B. R., Blaske, D. M., & Williams, R. A. (1995). Multisystemic treatment of serious juvenile offenders: Long-term prevention of criminality and violence. *Journal of Consulting and Clinical Psychology, 63,* 569–578.

Brestan, E. V., & Eyberg, S. M. (1998). Effective psychosocial treatments of conduct-disordered children and adolescents: 29 years, 82 studies, and 5,272 kids. *Journal of Clinical Child Psychology, 27,* 180–189.

Burns, G. L., & Patterson, D. R. (1990). Conduct problem behaviors in a stratified random sample of children and adolescents: New standardization data on the Eyberg Child Behavior Inventory. *Psychological Assessment, 2,* 391–397.

Burns, G. L., Patterson, D. R., Nussbaum, B. R., & Parker, C. M. (1991). Disruptive behaviors in an outpatient pediatric setting: Additional standardization data on the Eyberg Child Behavior Inventory. *Psychological Assessment, 3,* 202–207.

Capaldi, D. M., & Clark, S. (1998). Prospective family predictors of aggression toward female partners for at-risk young men. *Developmental Psychology, 34,* 1175–1188.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7–18.

Dishion, T. J., & Patterson, G. R. (1992). Age effects in parent training outcome. *Behavior Therapy, 23,* 719–729.

Egeland, B., Pianta, R., & Ogawa, J. (1996). Early behavior problems: Pathways to mental disorders in adolescence. *Development and Psychopathology, 8,* 735–749.

Eisenstadt, T. H., Eyberg, S., McNeil, C. B., Newcomb, K., & Funderburk, B. (1993). Parent-child interaction therapy with behavior problem children: Relative effectiveness of two stages and overall treatment outcome. *Journal of Clinical Child Psychology, 22,* 42–51.

Esser, G., Schmidt, M. H., & Woerner, W. (1990). Epidemiology and course of psychiatric disorders in school-age children—Results of a longitudinal study. *Journal of Child Psychology and Psychiatry, 31,* 243–263.

Eyberg, S. M., Boggs, S., & Algina, J. (1995). Parent-child interaction therapy: A psychosocial model for the treatment of young children with conduct problem behavior and their families. *Psychopharmacology Bulletin, 31,* 83–91.

Eyberg, S. M., & Ross, A. W. (1978). Assessment of child behavior problems: The validation of a new inventory. *Journal of Clinical Psychology, 16,* 113–116.

Hanf, C. (1969, April). *A two stage program for modifying maternal controlling during mother-child interaction*. Paper presented at the meeting of the Western Psychological Association, Vancouver, British Columbia, Canada.

Henggeler, S. W., Melton, G. B., & Smith, L. A. (1992). Family preservation using multisystemic therapy: An effective alternative to incarcerating serious juvenile offenders. *Journal of Consulting and Clinical Psychology, 60,* 953–961.

Henggeler, S. W., Rodick, J. D., Borduin, C. M., Hanson, C. L., Watson, S. M., & Urey, J. R. (1986). Multisystemic treatment of juvenile offenders: Effects on adolescent behavior and family interaction. *Developmental Psychology, 22,* 132–141.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336–352.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 559,* 12–19.

Kazdin, A. E., Bass, D., Siegal, T., & Thomas, C. (1989). Cognitive-behavioral therapy in the treatment of children referred for antisocial behavior. *Journal of Consulting and Clinical Psychology, 57,* 522–535.

Kazdin, A. E., Esveldt-Dawson, K., French, N. H., & Unis, A. S. (1987a). Effects of parent management training and problem-solving skills training combined in the treatment of antisocial child behavior. *Journal of the American Academy of Child and Adolescent Psychiatry, 26,* 416–424.

Kazdin, A. E., Esveldt-Dawson, K., French, N. H., & Unis, A. S. (1987b). Problem-solving skills training and relationship therapy in the treatment of antisocial child behavior. *Journal of Consulting and Clinical Psychology, 55,* 76–85.

Kazdin, A. E., Siegel, T. C., & Bass, D. (1992). Cognitive problem-solving skills training and parent management training in the treatment of antisocial behavior in children. *Journal of Consulting and Clinical Psychology, 60,* 733–747.

Kazdin, A. E., & Weisz, J. R. (1998). Identifying and developing empirically supported child and adolescent treatments. *Journal of Consulting and Clinical Psychology, 66,* 19–36.

Kendall, P. C., Flannery-Schroeder, E. C., & Ford, J. D. (1999). Therapy outcome research methods. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology.* New York: Wiley.

Kendall, P. C., & Grove, W. M. (1988). Normative comparisons in therapy outcome. *Behavioral Assessment, 10,* 147–158.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 285–299.

Lochman, J. E., Burch, P. R., Curry, J. F., & Lampron, L. B. (1984). Treatment and generalization effects of cognitive-behavioral and goal-setting interventions with aggressive boys. *Journal of Consulting and Clinical Psychology, 52,* 915–916.

McMahon, R. J. (1994). Diagnosis, assessment, and treatment of externalizing problems in children: The role of longitudinal data. *Journal of Consulting and Clinical Psychology, 62,* 901–917.

McNeil, C. B., Eyberg, S., Eisenstadt, T. H., Newcomb, K., & Funderburk, B. (1991). Parent-child interaction therapy with behavior problem children: Generalization of treatment effects to the school setting. *Journal of Clinical Child Psychology, 20,* 140–151.

Moffitt, T. E., Caspi, A., Dickson, N., Silva, P., & Stanton, W. (1996). Childhood-onset versus adolescent-onset antisocial conduct problems in males: Natural history from ages 3 to 18. *Development and Psychopathology, 8,* 399–424.

Patterson, G. R. (1976). *Living with children: New methods for parents and teachers*. Champaign, IL: Research Press.

Patterson, G. R., Forgatch, M. S., Yoerger, K. L., & Stoolmiller, M. (1998). Variables that initiate and maintain an early-onset trajectory for juvenile offending. *Development and Psychopathology, 10,* 531–547.

Pope, A. W., & Bierman, K. L. (1999). Predicting adolescent peer problems and antisocial activities: The relative roles of aggression and dysregulation. *Developmental Psychopathology, 35,* 335–346.

Reid, J. B. (1993). Prevention of conduct disorder before and after school entry: Relating interventions to developmental findings. *Development and Psychopathology, 5,* 243–262.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113,* 553–565.

Schuhmann, E. M., Foote, R. C., Eyberg, S. M., Boggs, S. R., & Algina, J. (1998). Efficacy of parent-child interaction therapy: Interim report of a randomized trial with short-term maintenance. *Journal of Clinical Child Psychology, 27,* 34–45.

Southam-Gerow, M. A., & Kendall, P. C. (1997). Parent-focused and cognitive-behavioral treatments of antisocial youth. In D. M. Stoff, J. Breiling, & J. D. Maser (Eds.), *Handbook of Antisocial Behavior* (pp. 384–394). New York: Wiley.

Spaccarelli, S., Cotler, S., & Penman, D. (1992). Problem-solving skills training as a supplement to behavioral parent training. *Cognitive Therapy and Research, 16,* 1–18.

Stattin, H., & Magnusson, D. (1996). Antisocial development: A holistic approach. *Development and Psychopathology, 8,* 617–645.

Taylor, T. K., Schmidt, F., Pepler, D., & Hodgins, C. (1998). A

comparison of eclectic treatment with parents and children series in a children's mental health center: A randomized clinical trial. *Behavior Therapy, 29,* 221–240.

Vuchinich, S., Bank, L., & Patterson, G. R. (1992). Parenting, peers, and the stability of antisocial behavior in preadolescent boys. *Developmental Psychology, 28,* 510–521.

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.

Webster-Stratton, C. (1990). Enhancing the effectiveness of self-administered videotape parent training for families with conduct-problem children. *Journal of Abnormal Child Psychology, 18,* 479–492.

Webster-Stratton, C. (1994). Advancing videotape parent training: A comparison study. *Journal of Consulting and Clinical Psychology, 62,* 583–593.

Webster-Stratton, C., & Hammond, M. (1997). Treating children with early-onset conduct problems: A comparison of child and parent training interventions. *Journal of Consulting and Clinical Psychology, 65,* 93–109.

Webster-Stratton, C., & Herbert, M. (1993). What really happens in parent training? *Behavior Modification, 17,* 407–456.

Webster-Stratton, C., Kolpacoff, M., & Hollinsworth, T. (1988). Self-administered videotape therapy for families with conduct-problem children: Comparison with two cost-effective treatments and a control group. *Journal of Consulting and Clinical Psychology, 56,* 558–566.

Received April 21, 2000; revised September 15, 2000; accepted November 22, 2000.