International Journal of Conflict and Violence

urn:nbn:de:0070-ijcv-2012239
IJCV: Vol. 6 (2) 2012, pp. 187 – 200

Clinical Significance of Parent Training for Children with Conduct Problems

Martin Forster, Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
Åsa Kling, Department of Psychology, Uppsala University, Sweden
Knut Sundell, The Swedish National Board of Health and Welfare (Socialstyrelsen), Stockholm, Sweden

Vol. 6 (2) 2012

ISSN: 1864–1385

Clinical Significance of Parent Training for Children with Conduct Problems

Martin Forster, Division of Psychology, Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
Åsa Kling, Department of Psychology, Uppsala University, Sweden
Knut Sundell, The Swedish National Board of Health and Welfare (Socialstyrelsen), Stockholm, Sweden

While there is a strong evidence base for behavioral parent training in the treatment of child conduct problems, the clinical impact is less well known. Meta-analyses report effect sizes in the medium range, but the common practice of reporting "small," "medium," and "large" effects can be misleading and difficult to understand for practitioners and clients. There is a need for more research addressing the clinical significance of behavioral parent training, which would help to bridge the gap between research and practice. In the first part of this report, a reanalysis in terms of clinical significance of two outcome studies published by the authors was conducted. In the second part, the results from the first part were compared to six outcome studies published by other authors. The median number needed to treat across studies was five, which means that for every five treated children, one shows reliable change and moves from the dysfunctional to the functional population.

Evidence for the efficacy of psychological interventions generally relies on reports of statistical and practical significance (e.g., Chambless and Hollon 1998). Although statistical significance testing provides information as to the reliability of outcomes, it tells us little about the importance of such outcomes. Practical significance (i.e., effect sizes) provides information as to the magnitude of treatment effects at a group level and has the advantage that effects can be compared across studies. It is however not easily understood by clinicians and can be influenced by factors such as within-group variance and baseline levels of outcome measures. Two studies with equal effect sizes can, for example, differ considerably in proportions of participants who recover or improve. This points to the inherent problem in using the conventional definitions of "small" (.20), "medium" (.50), and "large" (.80) effect sizes to classify treatment effects (Cohen 1988). A small effect size may be clinically meaningful in one context, while close to meaningless in others. Therefore, several researchers have stressed the importance of including clinical significance, in addition to statistical and practical significance, in reports of treatment effects (Campbell 2005; La Greca 2005). The primary aim of the current work was to investigate the clinical significance of behavioral parent training in the treatment of child conduct problems by synthesizing the results of published outcome studies.

Clinical significance refers to the importance or practical meaning of treatment effects – that is, proportions of clients who recover or improve and whether the changes make a real difference in the everyday life of the clients, besides reduction of the specific clinical symptoms being measured (Kazdin 1999). Despite the apparent benefits of considering clinical significance in syntheses of intervention research, it is seldom included in research reviews and clinical guidelines. The term was, for example, mentioned only once in an entire special issue of the *Journal of Clinical Child and Adolescent Psychology* on empirically supported treatments for children (initiated by American Psychological Association) (Silverman and Hinshaw 2008). There are at least two reasons for this. First, even if influential scholars and journals have called for the inclusion of clinical significance in outcome studies, most studies still only report results in terms of statistical significance and effect sizes (Ogles, Lunnen, and Bonesteel 2001; Campbell 2005). Second, there is no consensus as to how clinical significance

should be operationalized and measured (Ogles et al. 2001; Campbell 2005). Before to giving clinical significance the important role it deserves and including it in the guidelines for establishing empirically supported treatments as effective, it is essential to first agree upon its common operationalization and the analytical approach to its assessment.

The closest there is to a common standard is the procedure described by Jacobson and Truax (1991) – the "JT method" – which is the most widespread standardized method for assessment of clinical significance (Ogles et al. 2001). The JT method is based on two criteria that are used to classify participants in outcome studies. To satisfy the first criterion, an individual identified as member of the dysfunctional distribution on a given outcome measure must move to the functional distribution after treatment. This establishes *clinical change*. To satisfy the second criterion, the change in the individual has to be of sufficient magnitude to determine that it is significant rather than simply an artifact of measurement error. This establishes *reliable change*. Participants who satisfy both criteria are classified as *recovered*, whereas those who experience reliable change without passing the clinical cutoff are classified as *improved*.

The JT method has been shown to be as reliable and valid as more advanced statistical methods used to assess clinical significance (e.g., Bauer, Lambert, and Nielsen 2004). Therefore, and because the it is relatively easy to apply and understand, it is generally recommended over other methods (Bauer et al. 2004). It is important to point out that the JT method is not applicable to every type of clinical problem or context (Campbell 2005). The method does for example require that there is a clinically relevant cutoff point between dysfunctional (e.g., a diagnosis or defined risk group) and functional populations. In studies of problems without clinically relevant cutoff points (e.g., cigarette smoking), effect sizes may be a better way to operationalize meaningful change. On the other hand, for treatments targeting clinically defined groups, such as children with conduct problems, there is seldom an excuse not to use some variation of the JT method.

A strong argument for including clinical significance in research reports is that policymakers, practitioners, and consumers can more easily understand the magnitude of treatment effects. Some authors also argue that results obtained in analyses of clinical significance should preferably be reported as numbers needed to treat – NNT (Marrs-Garcia 2010). The NNT is the number of individuals who would need to be exposed to a particular treatment before one individual would recover. Hence, a NNT close to 1 suggests that nearly all study participants recovered. In controlled studies, the NNT represents the relative advantage of the treatment group over the control group. For example, if every other participant (50 percent) recovered in a treatment group and every fourth (25 percent) spontaneously recovered in a no-treatment control group, the "net" gain of the treatment is 25 percent (50 percent minus 25 percent). For every four treated patients one would recover as a result of the treatment, which translates to a NNT of four.

Like intervention research in general, research on behavioral parent training for children with conduct problems suffers from a lack of standardized analyses of clinical significance. In most outcome studies of behavioral parent training reports of clinical significance are not included at all. Some studies use procedures that prevent comparison across studies, such as defining clinical significance as participants who show at least 30 percent improvement on a given outcome measure (e.g., Reid, Webster-Stratton, and Hammond 2007). Only a handful of studies use standardized methods (e.g., the JT method) that allow for comparison across studies and synthesis of data. The omission of clinical significance in published reviews and meta-analyses of behavioral parent training in the treatment of child conduct problems therefore comes as no surprise. This stands in sharp contrast to the impressive body of research supporting the statistical and practical significance of behavioral parent training, which has been reported in numerous reviews and meta-analyses (Eyberg, Nelson, and Boggs 2008; Furlong et al. 2012; Dretzke et al. 2009).

In the first part of this article, the aim is to contribute to the small body of research that properly reports clinical significance in studies of behavioral parent training for children with conduct problems by reanalyzing two studies previously published by some of the authors of the present

report in terms of clinical significance. In the second part, we investigate the clinical significance of behavioral parent training by synthesizing the results from the handful of published studies that have used standardized procedures to assess the clinical significance of behavioral parent training.

1. Part I: Reanalysis of Two Published Parent Training Studies

In the first part of this contribution, we report an analysis of clinical significance performed with data from two previously published studies by some of the authors of the present report. Both studies were randomized trials of a Swedish parent training program called Comet (Kling et al. 2010; Enebrink et al. 2012). In both studies there were statistically significant differences between treatment groups and waitlist control groups, with effect sizes in the medium to large range. The program has been implemented on a wide scale in Sweden through different methods of delivery. The standard method of delivery (Comet-S) consists of eleven 2.5-hour workshops, in which two practitioners, usually from the social services, provide guidance in effective parenting practices to groups of parents. The program is based on a manual, which contains theory and practice in sensitive play, praise, incentives, ignoring of misconduct, and rules and expectations. Video modeling, role-play, and homework assignments are key ingredients in the process of delivery. Parents participating in the self-directed version of the program (Comet-SD) receive exactly the same written material as parents in Comet-S, but the material is introduced at a single workshop without further practitioner support. The internet-based delivery format (Comet-I) also contains the same material as Comet-S, including instructional text and video vignettes, but also offers several interactive features such as participant support forums and minimal e-mail contact with a practitioner. For further description of the content and evaluations of the Comet program, see Kling et al. (2010) and Enebrink et al. (2012).

1.1. Method

1.1.1. Analysis of Clinical Change

The JT method was used in the reanalysis of the two studies. The first step of the JT method is to determine whether participants experience a clinical change, i.e., move from the dysfunctional to the functional distribution on a given outcome measure. To make such an analysis, a cutoff point that divides the two distributions has to be determined. Cutoff C, which is defined as the weighted midpoint between the means of functional and dysfunctional populations, is generally the recommended method (Bauer et al. 2004; Evans, Margison, and Barkham 1998). Computation of Cutoff C requires that normative data is available for the selected outcome measure(s). While several outcome measures were used in the studies, norms were only available for the Eyberg Child Behavior Inventory (ECBI). Therefore, that measure was used to assess clinical significance in the two studies.

The ECBI (Eyberg and Pincus 1999) consists of thirty-six items describing disruptive and aggressive behaviors (e.g., "Hits parents" and "Does not obey house rules"), which are each rated in terms of their frequency on a seven-point likert scale (1 = never happens, 7 = always happens). The sum of these items is called the "intensity scale" (ECBI-IS) with a range of 36–252. The same items are also rated on a "problem scale" (ECBI-PS), which measures whether the parents experience the occurring behaviors as problematic (1 = yes) or not (0 = no). That scale thus has a range of 0–36. The ECBI is probably the most common outcome measure in studies of behavioral parent training for children with conduct problems (Dretzke et al. 2009) and numerous studies have investigated and confirmed its psychometric properties (Plake, Impara, and Spies 2003). In the study by Kling et al. (2010), mothers alone were the respondents for 84 percent of the participants. Fathers alone were respondents for 10 percent of the participants and both parents responded for the final 6 percent. In the study by Enebrink et al. (2012), it is unknown who the responding parent was.

The means and standard deviations for the dysfunctional population (i.e., children with conduct problems) and functional population (i.e., normal children) are required to compute the cutoff C. The pretest means and standard deviations in each study were used to represent the dysfunctional population in the present analysis, which is recommended instead of using published normative/clinical data (Jacobson and Truax 1991). On the other hand, published normative data is required to obtain means and

standard deviations for functional populations. In the present analysis, Swedish normative data was used as means and standard distributions for the functional distributions (Axberg, Johansson Hanse, and Broberg 2008). Normative data for six-year olds were used for the analysis of Kling et al. (2010); $M = 90.7$, $SD = 23.6$ for ECBI-IS and $M = 2.95$, $SD = 4.10$ for ECBI-PS. For Enebrink et al. (2012), normative data for seven-year olds was used; $M = 85.2$, $SD = 23.5$ for ECBI-IS and $M = 2.46$, $SD = 4.08$ for ECBI-PS. The weighted midpoint (Cutoff C) between functional and dysfunctional distributions was 113 (ECBI-IS) and 8 (ECBI-PS) for participants in Kling et al. (2010). In Enebrink et al. (2012), the corresponding cutoff points were 122 (ECBI-IS) and 9 (ECBI-PS). Participants who scored above these cutoff points at pretest, and below at posttest, satisfied the criterion for clinical change.

1.1.2. Analysis of Reliable Change

In the second step of the JT method, a reliable change index (RCI) is computed for each participant, representing the change between pretest and posttest divided by the standard error of difference between the two scores (Jacobson and Truax 1991). The standard error of difference is dependent on the variability in the studied sample (i.e., the standard deviation at pretest), but also the reliability of the measurement. The internal consistency of the ECBI, which is recommended over other types of reliability measures, was used as the reliability coefficient in the present analyses (Bauer et al. 2004; Evans et al. 1998). Furthermore, the reliability coefficient should be obtained from the studied sample, rather than published test data (Campbell 2005). In Kling et al. (2010), the internal consistency (Cronbach's alpha) was $\alpha = .92$ (ECBI-IS) and $\alpha = .89$ (ECBI-PS), while the corresponding coefficients were $\alpha = .81$ (ECBI-IS) and $\alpha = .79$ in Enebrink et al. (2012).

For individuals with a reliable change index larger than 1.96 change is unlikely to be due to measurement error ($p < .05$), which means that they satisfy the criterion for reliable change. It is also possible to calculate how much an individual must change on a given outcome. For participants in Kling et al. (2010), the minimum difference between pretest and posttest that constituted a reliable change was 20.4 points on the ECBI-IS and 5.6 points on the ECBI-PS. The corresponding thresholds for participants in Enebrink et al. (2012) were 22.4 (ECBI-IS) and 7.0 (ECBI-PS).

1.1.3. Classification of Participants

The participants in the current analysis were classified as *recovered* if they made both a reliable and a clinical change (satisfied both criteria in the JT method). They were classified as *improved* if they satisfied the criterion of reliable change, but not that of clinical change. If they made a reliable change in the undesired direction, they were classified as *deteriorated*. Finally, participants who made no reliable change in any direction were classified as *unchanged*. Sometimes the unchanged category is defined as participants who "pass neither criteria" (e.g., Campbell, 2005), but McGlinchey, Atkins, and Jacobson (2002) recommend the definition used here. Finally, chi-square analyses (Fisher's exact test) were performed to assess whether the clinical significance differed significantly between the treatment and control groups, as recommended by Kendall et al. (1999).

Using the intention to treat principle in the analyses of the two studies makes particular sense with regard to clinical significance. Early termination of treatment may be en even bigger problem in clinical practice than in research settings (Kazdin 2008) and it is therefore reasonable to assess clinical significance including the total sample rather than just study completers. To obtain complete data for every participant, the last observed score was carried forward in cases of missing data at posttest and/or follow-up. This implies that every participant who dropped out or had a missing score was classified as unchanged.

1.2. Results

Table 1 shows the effect sizes and clinical significance for the three study conditions in Kling et al. (2010). After subtracting the proportions of the control group from the treatment groups, the recovery rates for the ECBI-IS were 28 percent (Comet-S), 13 percent (Comet-SD), and 26 percent (Comet-I), which translates to NNTs of four, eight, and four. The corresponding rates for the ECBI-PS were 28 percent, 11 percent, and 39 percent, with NNTs of four, nine, and 13. The recovery rates were statistically significantly larger in all treatment groups compared to the waitlist control groups.

Table 1: Clinical significance of the Comet program at posttest

Program/outcome	Cohen's *d*	Recovered n (%)	Improved n (%)	Unchanged n (%)	Deteriorated n (%)
Comet-S					
ECBI-IS	.71	16 (28)***	6 (10)	35 (60)*	1 (2)
ECBI-PS	.90	16 (28)***	8 (14)	34 (58)*	1 (2)
Comet-SD					
ECBI-IS	.56	8 (13)*	7 (11)	45 (74)	1 (2)
ECBI-PS	.52	7 (11)*	7 (11)	45 (74)	2 (4)
Waitlist					
ECBI-IS	.01	0 (0)	4 (10)	33 (82)	3 (8)
ECBI-PS	.00	0 (0)	3 (7)	32 (80)	5 (13)
Comet-I					
ECBI-IS	1.62	30 (52)*	3 (5)	20 (34)**	5 (9)
ECBI-PS	1.53	30 (52)***	6 (10)	21 (36)***	1 (2)
Waitlist					
ECBI-IS	.83	12 (26)	3 (7)	31 (67)	0 (0)
ECBI-PS	.72	6 (13)	3 (7)	37 (80)	0 (0)

Notes: Cohen's *d*s are within-group effect sizes.
Proportion significantly different from corresponding proportion in waitlist control group at * p < .05, ** p < .01, *** p < .001.

Table 2 shows the results at six-month follow-up for Comet-S and Comet-SD. The proportions of recovered or improved participants were larger or similar to the corresponding proportions at posttest in both groups. The recovery rates were about twice as large for Comet-S (29 percent and 43 percent) as for Comet-SD (15 percent and 18 percent), but only the advantage pertaining to ECBI-PS was statistically significant. No follow-up data on clinical significance for Comet-I is reported in Enebrink et al. (2012).

Table 2: Clinical significance of Comet-S versus Comet-SD at follow-up

Program/outcome	Cohen's *d*[a]	Recovered n (%)	Improved n (%)	Unchanged n (%)	Deteriorated n (%)
Comet-S					
ECBI-IS	.85	17 (29)	9 (16)	30 (52)	2 (3)
ECBI-PS	1.20	25 (43)**	8 (14)	24 (41)*	1 (2)
Comet-SD					
ECBI-IS	.89	9 (15)	15 (25)	36 (59)	1 (2)
ECBI-PS	.82	11 (18)	10 (16)	37 (61)	3 (5)

Notes: Cohen's *d*s are within-group effect sizes (pretest/follow-up).
Proportion significantly different from corresponding proportion in Comet-SD group at * p < .05,; ** p < .01.

2. Part II: Synthesis of Results across Studies

2.1. Method

2.1.1. Inclusion and Exclusion of Studies

The second part of this contribution compares published studies of behavioral parent training that include reports of clinical significance. The databases of PsychInfo and PubMed were searched up to July 2012. In addition, citations from a recent meta-analysis of behavioral parent training (Dretzke 2009) were also investigated. Marrs-Garcia (2010) specifies three conditions that have to be fulfilled to enable meaningful comparisons of NNTs across studies, which also apply to comparisons of clinical significance in general: (a) clinical significance has to be operationalized the same way across studies, (b) the control or comparison groups to which treated groups were compared have to be equivalent, and (c) the same outcome measure has to be used across studies. With these guidelines in mind, a set of criteria was developed to select studies for inclusion. First, only studies that based the analysis of clinical significance on the JT method, including analysis of both reliable and clinical change, were included. Second, only studies that included a waitlist/no-treatment control group were included, because this was the only type of comparison group that occurred in several studies. Third, only studies that based the analysis of clinical significance on the ECBI were included, because that was the only measure that occurred with sufficient frequency to allow proper comparisons across studies. Fourth, only studies published in peer-reviewed journals were included.

The search found in twenty-one studies of behavioral parent training for children with conduct problems that compared the treatment to a waitlist/no-treatment control group and employed the JT method to assess clinical significance. Five studies were excluded for reporting only reliable, but not clinical change, and another five studies were excluded for the opposite reason. Finally, three studies were excluded for basing the analysis of clinical significance on measures other than the ECBI. No authors of the excluded studies were contacted, because it was considered difficult or impossible for them to perform the necessary analyses to make the studies eligible for inclusion. The final sample therefore consisted of eight studies (including the two from the first part of this report), altogether including 13 treatment conditions (Table 3.

In four of the studies in Table 3, different versions of the Triple-P program (Sanders 1999) were evaluated. Triple-P is a multilevel behavioral parent training program that targets different risk groups of children with conduct problems. In the self-directed version of the program (Triple-P-SD), parents receive training material (video and text) that they implement without any practitioner support. Triple-P-SD has also been enhanced in some studies with limited telephone support and/or a single session led by practitioners (Triple-P-SD+). In the standard version of the program (Triple-P-S), parents take part in ten individual one-hour sessions with a practitioner. Finally, the program has also been offered as an enhanced version (Triple-P-E). In addition to the ten sessions offered in Triple-P-S, parents in Triple-P-E also receive ten to fifteen sessions involving strategies to increase support from partners and friends as well as methods to manage stress, anxiety, and depression.

In the study by Nixon et al. (2003), two versions of the Parent-Child Interaction Therapy (PCIT) (Eyberg 1988) were evaluated. In the standard version of the program (PCIT-S), parents take part in twelve sessions (one to two hours) with a practitioner. In an abbreviated version of the program (PCIT-ABB), parents receive videotapes to learn the skills that are taught in PCIT-S. They also attend five face-to-face sessions with a practitioner, alternated with brief telephone sessions.

In the last study in Table 3, The Incredible Years program (IY) (Webster-Stratton 2000) was evaluated. In that program, parents of six to eight children meet for twelve to fourteen weekly two-hour sessions. Several video vignettes on specific parenting skills are shown and discussed during the sessions. Skills are role-played in the group and the parents get a weekly assignment to practice their newly acquired skills at home between sessions.

In all studies, with the exception of Enebrink et al. (2012), it was possible to conclude that the analysis of clinical significance was mainly or completely based on mothers' responses on the ECBI.

Table 3: Characteristics of synthesized studies

Study	Study conditions	n	Child age *M (SD)*	ECBI-IS pretest *M (SD)*	Cutoff
Kling et al. (2010)	Comet-S	58	6.0 (2.3)	138.0 (26.0)	113[a]
	Comet-SD	61			
	Waitlist	40			
Enebrink et al. (2012)	Comet-I	58	6.8 (2.3)	150.7 (18.5)	122[a]
	Waitlist	46			
Sanders et al. (2000)	TripleP-E	76	3.4 (0.3)	152.8 (26.0)	Not specified
	TripleP-S	77			
	TripleP-SD	75			
	Waitlist	77			
Morawska and Sanders (2006)	TripleP-SD+	43	2.2 (0.4)	119.1 (26.4)	131[b]
	TripleP-SD	42			
	Waitlist	41			
Morawska et al. (2011)	TripleP-ABB	33	3.6 (0.9)	146.6 (28.0)	131[b]
	Waitlist	34			
Joachim, Sanders, & Turner (2010)	TripleP-ABB	26	3.3 (1.1)	129.4 (25.8)	131[b]
	Waitlist	20			
Nixon et al. (2003)	PCIT-S	22	3.9 (0.6)	164.9 (19.4)	131[b]
	PCIT-ABB	23			
	Waitlist	18			
Axberg and Broberg (2012)	IY	38	6.0 (1.3)	156.4 (21.4)	121[b]
	Waitlist	24			

[a] The cut-off point was the weighted midpoint between the study sample mean and the mean of a normative population, as recommended by Jacobson and Truax (1991).
[b] Cut-off point based on normative data only.

2.1.2. Analytic Strategy

Several different outcomes were included in the comparison of clinical significance across studies. First, the effect sizes were computed (Cohen's *d*), to enable comparison between practical and clinical significance. To make the effect sizes comparable across studies, they were not retrieved from the original articles, but re-computed from reported means and standard deviations. First, within-group effect sizes were computed separately for the treatment and control groups. The pooled standard deviation at pretest was used as denominator, with correction for small samples, and pre-post change scores were used as numerator. Second, the between-group effect sizes were computed by subtracting the within-group effect size in each control group from the corresponding treatment group.

Second, the proportions of participants experiencing reliable and clinical change were computed. When results are to be compared across studies, it is necessary to analyze between-group effects (Marrs-Garcia 2010), which therefore were computed by subtracting proportions of reliable and clinical change in the control groups from the corresponding proportions in the intervention groups. For example, in Kling et al. (2010), 38 percent of the participants in Comet-S and 10

percent in the waitlist control group experienced reliable change. The reliable change in terms of between-group effects therefore was 28 percent (38 minus 10).

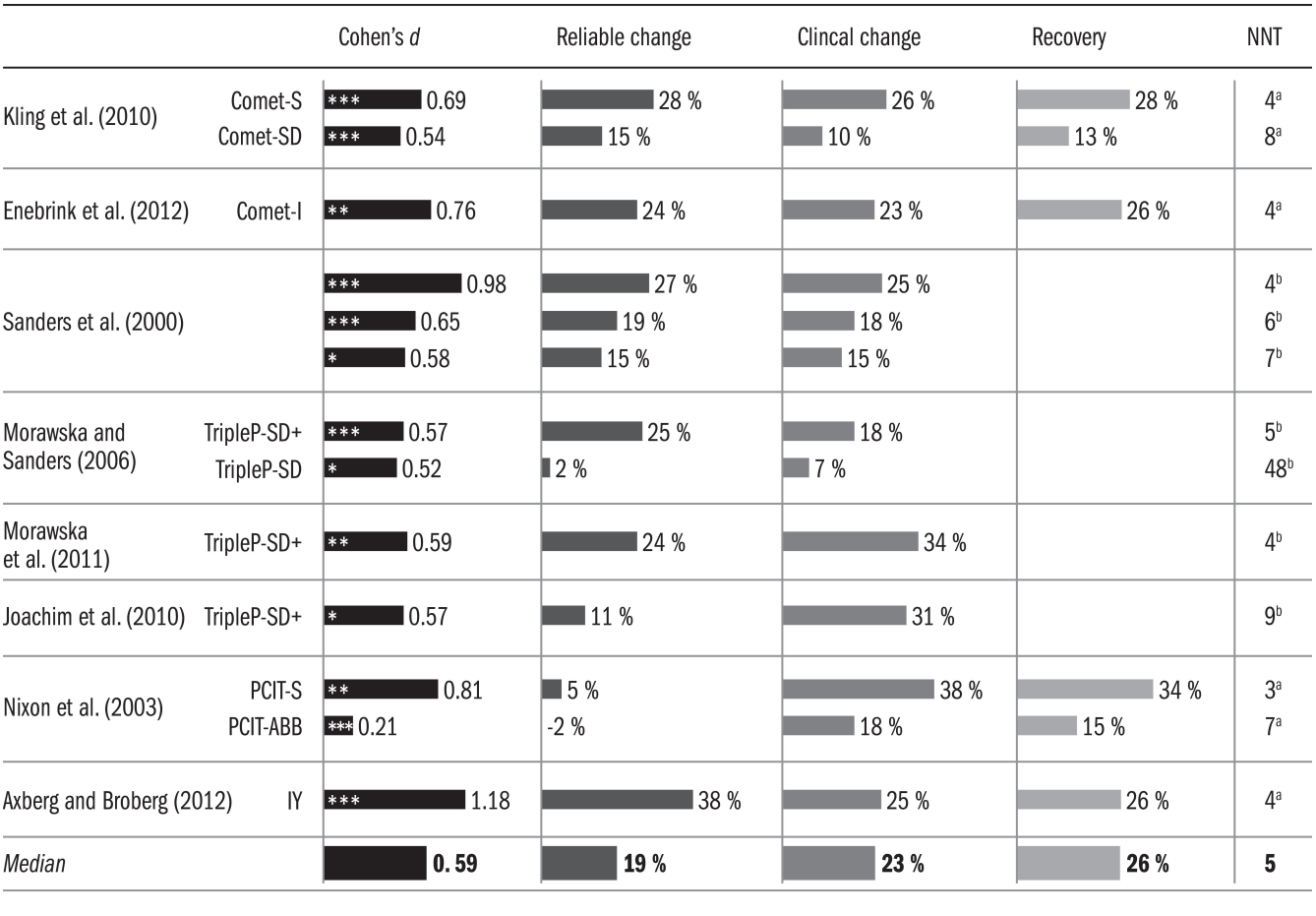
Third, NNTs were operationalized and computed in two different ways. Some of the included studies used the classification of participants, as suggested in the JT method (Enebrink et al. 2012; Kling et al. 2010; Nixon et al. 2003). For those studies, NNTs based on proportions of recovered participants were computed. In the other studies, the JT method was used to compute proportions of participants experiencing reliable and clinical change, but the proportions were not combined to classify participants as recovered, improved, unchanged, or deteriorated, as suggested by Jacobson and Truax (1991). For those studies, the NNTs were based on either the reliable change or clinical change, whichever proportion was the smallest.

In several of the included studies the reported clinical significance was based on participants who completed the study, with no account of dropouts . In this synthesis, all results were instead analyzed as intention to treat. Dropouts were consequently counted as unchanged, which corresponds to the last observation carried forward method of handling missing data. A few studies also included follow-up measurements, but the dropout rates were generally high. Therefore, this report only include results at posttest in the synthesis across studies (Figure 1).

2.2. Results

The effect size and clinical significance based on the ECBI-IS for each of the thirteen treatments from the eight included studies are presented in Figure 1. The median was computed, instead of the mean, due to large variability across studies and treatments. Most effect sizes were in the medium to large range and all were of sufficient magnitude to be statistically significant. The median effect size across the thirteen treatments was $d = 1.31$ based on within-group effect sizes, and $d = .59$ based on between-group effect sizes (treatment vs. control). The median proportion of participants who made reliable change was 38 percent (within-group) and 19 percent (between-group). The corresponding proportions for clinical change were 37 percent (within-group) and 23 percent (between-group). For six treatments in Figure 1, proper recovery rates combining the criteria of reliable and clinical change were available. Across treatments, the median proportion of participants who experienced recovery was 31 percent (within-group) and 26 percent (between-group). Because the NNT represents a comparison between treatment and control conditions in controlled studies, all NNTs represent between-group effects. The median NNT across all treatments was five, which means that for every five treated children, one recovers. The median NNT for the treatments that involved full practitioner support (Comet S, TripleP-E, TripleP-S, PCIT-S, and IY) was four, while the median for the other treatments that involved no or minimal support was seven. Five of the included studies also reported the clinical significance based on the ECBI-PS (Kling et al. 2010; Enebrink et al. 2012; Morawska and Sanders 2006; Morawska et al. 2011, Joachim et al. 2010). Compared to the ECBI-IS, the NNTs based on the ECBI-PS were similar or slightly lower, with a median across treatments of four.

Figure 1: Effect size, reliable change, clinical change, and NNT at posttest (based on ECBI-IS)



Notes:
All results are between-group effects (i.e., the effects or proportions of the waitlist control groups are subtracted from those for the treatment groups). The significance levels of the effect sizes (*p < .05, **p < .01, ***p < .001) were retrieved from the original articles. NNT = Numbers Needed to Treat.
[a] The NNT is based on the proportion of participants who recovered according to the JT method.
[b] The NNT is based on whichever of the reliable or clinical change proportions was smallest.

3. Discussion

Part I investigated the clinical significance of behavioral parent training for children in two studies previously published by the authors. In Kling et al. (2010), about one fourth of the participants recovered in the practitioner-assisted version of the program (Comet-S), while only about one participant in eight made a recovery in the self-directed version of the program (Comet-SD). The relative advantage of Comet-S over Comet-SD was preserved at follow-up. This result is worth noting, considering that the advantage of Comet-S was less apparent in the original article where results were reported in terms of statistical and practical significance (Kling et al. 2010). One interpretation of this result is that practitioners played an important role in helping clients making an actual recovery. While many participants in the self-directed version did improve, as shown by the effect size, most of the changes at the individual level were too small to be clinically significant.

In Enebrink et al. (2012) the within-group recovery rate in the treatment group was considerably higher than in Kling et al. (2010). However, after taking the control group into account, the recovery rates were similar to Comet-S. It is still striking that an internet-based version of the program with minimal practitioner assistance (Comet-I) was as effective as Comet-S in terms of clinical significance. One possible bias may be that the samples in the two studies were quite different. For example, only one third of the participating parents in the Comet-S study were well educated, compared to two thirds in the Comet-I study. It is well known that the social characteristics of families that take part in behavioral parent training can impact the treatment effects (Reyno and McGrath 2006). It is also worth noting that almost one in every ten parents in the Comet-I condition deteriorated. Even if this number was non-significant compared to the waitlist, it may warrant further investigation.

Part II investigated the clinical significance of behavioral parent training for children with conduct problems by synthesizing results from published studies. The median NNT was five across all treatments, four for the treatments offering full practitioner support, and seven for treatments offering no or minimal practitioner support. Even in the most effective programs, the NNTs were not lower than three. This means that, at best, one third of children with conduct problems actually recover as a result of treatment with behavioral parent training, while the rest only improve to some degree, show no change, or even deteriorate. This result is important for at least two reasons. First, the between-group recovery rates presented in Figure 1 are considerably lower than several of the within-group rates reported in the original articles. We argue that the former rates are of greater interest to practitioners and clients, because they account for bias such as spontaneous recovery and instead reflect the proportion of participants who recover *as a result of* the treatment. Second, many practitioners and clients may not realize that programs that are characterized as evidence-based and have reported "large" effect sizes, will only "cure" a minority of the treated children. However, from a researcher or policymaker point of view, curing one out of three patients may be of tremendous importance. The meaning of treatment effects is context-dependent and factors such as severity of the treated problems and cost-benefit analyses have to be considered (Campbell 2005). For example, the similar effect sizes between Comet-S and Comet-SD at follow-up in Kling et al. (2010) mask the fact that the proportion of children making an actual recovery was much larger in Comet-S. The higher cost of implementing Comet-S compared to Comet-SD would be returned many times if twice as many children fully recovered from conduct problems.

The fact that only eight studies were included in the second part of this report limited the possibility to draw conclusions about the effects of moderators. It made little sense, for example, to compare the effects of different programs or the effect of child age, when there were only one or two studies representing a certain program or age group. It was however less problematic to investigate the effects of different methods of delivery on clinical significance, because such comparisons could be made both within and across studies. Therefore, method of delivery is the only moderator that can be discussed in any depth. The effects of different methods of delivery within the same trial were investigated in four studies (Kling et al. 2010; Morawska et al. 2006; Nixon et al. 2003; Sanders et al. 2000). As in Kling et al. (2010), the analyses of the other studies also showed that the clinical significance was larger for treatments involving more practitioner support (Figure 1). In some studies, advantages for treatment conditions involving more practitioner support were apparent in terms of clinical significance, but not in terms of statistical significance or effect sizes. For example, in Morawska and Sanders (2006) the effect sizes were similar for the two compared treatments, but there was a large difference between NNTs. The NNTs of self-directed treatments in Enebrink et al. (2012), Morawska et al. (2006) and Morawska et al. (2011) were similar to the most effective of programs offering full practitioner support in Figure 1. However, as in Enebrink et al. (2012), the participants in the two studies by Morawska and colleagues were particularly well educated and had few social problems. The other studies in Figure 1 all recruited average or at-risk samples. In conclusion, practitioner support seems to have a greater impact on clinical significance than on statistical significance and effect sizes. This conclusion at least holds based on comparisons of dif-

ferent levels of practitioner support within studies, which compared to comparisons across studies are less influenced by possible confounding variables such as characteristics of the study sample.

Two results that were found in the synthesis (Figure 1) especially warrant discussion. In Morawska et al. (2006), very few participants made a reliable and clinical change in the Triple-P-SD condition compared to the Triple-P-SD+, despite the fact that the effect sizes were similar in the two treatment groups. This means that many of the participants in Triple-P-SD must have improved, but not by enough to satisfy the criteria for reliable and clinical change. In Nixon et al. (2003), an unusually large proportion of participants in the waitlist control group made a reliable change (50 percent). The between-group proportions of reliable change thus turned out to be very small or even negative. Still, the NNTs were quite small for the treatment groups in that study. The reason for this was that they were based on recovery rates. Only 11 percent of the participants actually recovered in the control group, while 45 percent recovered in PCIT-S and 26 percent recovered in PCIT-ABB.

Several limitations of this report are recognized. First of all, the small number of studies included in the synthesis limited the possibility for more detailed analyses and generalization of the results. However, more liberal inclusion criteria would have made comparison across studies impossible (Marrs-Garcia 2010). It is therefore imperative that future studies of behavioral parent training include standardized analysis of clinical significance, preferably using the JT method (e.g., Bauer et al. 2004; McGlinchey et al. 2002). Another possibility would be to reanalyze the original data from a larger number of published outcome studies of behavioral parent training that lack reports of clinical significance.

A second limitation is that the analyses of clinical significance in this article were based on only one outcome measure. In Kling et al. (2010) and Sanders et al. (2000), a structured telephone interview measuring child conduct problems was also used as a basis for analyses of clinical significance (Parent Daily Report or PDR; Chamberlain and Reid 1987). Due to lack of proper normative data, the PDR could not be used to analyze clinical change. The reliable change proportions were, however, computed. Based on the PDR, the reliable change proportions were 6 percent for Comet-SD and 5 percent for Triple-P-SD, as compared to 15 percent in both programs when the analyses were based on the ECBI-IS. There are a number of possible explanations for this difference, which also apply to differences in reliable change across measures in general. It could reflect general characteristics of the measures, such as sensitivity to change, reactivity, and different forms of reliability and validity. Further, the parameters that are used to compute the reliable change index are obvious sources of variation. For example, the internal consistency was .92 for the ECBI-IS as compared to .79 for the PDR in Kling et al. (2010), which strongly impacts the resulting thresholds for reliable change. To conclude, the choice of measures will often have an impact on results in analyses of clinical significance. Instead of relying on single measures, a compound of measures that target the construct of interest should ideally be used in analyses of clinical significance (Ogles et al. 2001). This is, however, often impossible due to lack of published norms for many outcome measures. Furthermore, because clinical significance refers to meaningful changes in the everyday life of clients (Kazdin 1999), it would also often be necessary to base analyses of clinical significance on several outcome domains (e.g., quality of life), besides the treated symptoms (e.g., child conduct problems). It is however by no means certain that a narrow focus on measuring symptoms will overestimate the clinical significance in terms of everyday functioning, as there are examples of the opposite (Karpenko et al. 2009).

A third limitation pertains to the application of the JT method in the six studies located by database search (Axberg and Broberg 2012; Joachim et al. 2010; Morawska et al. 2006; Morawska et al. 2011; Nixon et al. 2003; Sanders et al. 2000). First, none of the Triple-P-studies combined the criteria of reliable and clinical change. The discrepant results found in the analyses of Nixon et al. (2003) illustrate that it can be quite misleading to report reliable and clinical change separately, as opposed to combining the two criteria according to the JT method. This also justifies the exclusion of ten studies reporting only reliable or only clinical change in the process of finding eligible studies for the synthesis.

Second, none of the studies that were included from the database search reported which reliability coefficient were used to compute reliable change (e.g., internal consistency or test-retest). This limits the transparency and accuracy of the comparison of results across studies. Third, none of the six studies seem to have applied the recommended cutoff for clinical change, which is the weighted midpoint between the functional and dysfunctional populations (Jacobson and Truax 1991). The theory underlying the recommended cutoff point is that each study in essence investigates a unique population, whose cutoff in relation to the normative population also will be unique. It is generally difficult to compare results across studies with different populations, and using a cutoff point that partly is based on the study sample is probably more accurate than imposing an absolute cutoff (Wise 2004). With an absolute cutoff point, there is a risk that a substantial number of participants happen to score just above (or just below) the cutoff at pretest, which will result in misleading proportions for clinical change. This may have been the case in Joachim et al. (2010), where the pretest mean was close to the selected cutoff point (Table 3). Fourth, the potential for clinical change was limited in several studies because a substantial number of participants already scored below the cutoff at pretest. This was not the case in Enebrink et al. (2012) and Axberg and Broberg (2012), where only 5–10 percent of participants were below the cutoff at pretest, and probably not in Sanders et al. (2000), which used an elevated ECBI score as inclusion criterion. It was more of a problem in Kling et al. (2010) and Morawska et al. (2011) with 21 percent and 24 percent respectively scoring below the cutoff at pretest. In Joachim at al. (2010), 50 percent of the participants scored below the cutoff at pretest and in Morawska and Sanders (2006) the proportion was as large as 62 percent. Sometimes analyses of subsamples of participants who score above the cutoff at pretest are used in such cases, but in randomized trials that strategy can result in selection bias. Instead, if a study intends to investigate treatment effects for a defined population, proper screening should be used to ensure that included participants actually belong to the dysfunctional/clinical population. Alternatively, several cutoff points could be used to represent different levels of severity of a given problem or condition (Ogles et al. 2001). Such a procedure could also be warranted, given that some authors suggest that the JT method sometimes may be too conservative (Tingey et al. 1996).

4. Conclusion

The results of this synthesis shows that the effects of behavioral parent training in the treatment of child conduct problems are generally clinically significant, but maybe to a lesser degree than would have been expected. The median recovery rate across studies showed that only one out of five children recovered. It was also evident that results in terms of clinical significance may lead to different conclusions than where conclusions are based solely on results in terms of statistical significance and effect sizes. The effects of practitioner support were considerably stronger in terms of clinical significance than in terms of effect sizes. The results further support the importance of including reports of clinical significance in outcome studies, which has called for by influential scholars and journals (Ogles et al. 2001; Campbell 2005). The fact that only eight studies were found to be eligible for inclusion in this synthesis points to the need for future research to adopt similar standards for the analysis of clinical significance, such as the JT method. A study that states that one out of five patients recover, rather than saying that the effect was d = .59, would probably be more effective in bridging the gap between science and practice.

References

Axberg, Ulf, and Anders G. Broberg. 2012. Evaluation of the "Incredible Years" in Sweden: The Transferability of an American Parent-Training Program to Sweden. *Scandinavian Journal of Psychology* 53:224–32.

Axberg, Ulf, Jan Johansson Hanse, and Anders G. Broberg. 2008. Parents' Description of Conduct Problems in their Children – A Test of the Eyberg Child Behavior Inventory (ECBI) in a Swedish Sample Aged 3–10. *Scandinavian Journal of Psychology* 49:497–505.

Bauer, Stephanie, Michael J. Lambert, and Steven L. Nielsen. 2004. Clinical Significance Methods: A Comparison of Statistical Techniques. *Journal of Personality Assessment* 82:60–70.

Campbell, Todd C. 2005. An Introduction to Clinical Significance: An Alternative Index of Intervention Effects for Group Experimental Designs. *Journal of Early Intervention* 27:210–27.

Chamberlain, Patricia., and John B. Reid. 1987. Parent Observation and Report of Child Symptoms. *Behavioral Assessment* 9:97–109.

Chambless, Dianne L., and Steven D. Hollon. 1998. Defining Empirically Supported Therapies. *Journal of Consulting and Clinical Psychology* 66:7–18.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2d ed. Hillsdale, NJ: Erlbaum.

Dretzke, Janine, Clare Davenport, Emma Frew, Jane Barlow, Sarah Stewart-Brown, Sue Bayliss, Rod S. Taylor, Josie Sandercock, and Chris Hyde. 2009. The Clinical Effectiveness of Different Parenting Programmes for Children with Conduct Problems: A Systematic Review of Randomized Controlled Trails. *Child and Adolescent Psychiatry and Mental Health* 3:7.

Enebrink, Pia, Jens Högstrom, Martin Forster, and Ata Ghaderi. 2012. Internet-Based Parent Management Training: A Randomized Controlled Study. *Behaviour Resarch and Therapy* 50: 240–49.

Evans, Chris, Frank Margison, and Michael Barkham. 1998. The Contribution of Reliable and Clinically Significant Change Methods to Evidence-Based Mental Health. *Evidence-Based Mental Health* 1:70–72.

Eyberg, Sheila M. 1988. Parent–Child Interaction Therapy: Integration of Traditional and Behavioral Concerns. *Child and Family Behavior Therapy* 10:33–46.

Eyberg, Sheila M., Melanie M. Nelson, and Stephen R. Boggs. 2008. Evidence-Based Psychosocial Treatments for Children and Adolescents with Disruptive Behavior. *Journal of Clinical Child and Adolescent Psychology* 37:215–37.

Eyberg, Sheila M., and Donna Pincus. 1999. *Eyberg Child Behavior Inventory and Sutter-Eyberg Student Behavior Inventory – Revised, Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Furlong, Mairead., Sinead McGilloway, Tracey Bywater, Judy Hutchings, Susan M. Smith, and Michael Donnelly. 2012. Behavioural and Cognitive-Behavioural Group-Based Parenting Programmes for Early-Onset Conduct Problems in Children Aged 3 to 12 Years. *Cochrane Database of Systematic Reviews* 2 CD008225. doi:10.1002/14651858.CD008225.pub2

Jacobson, Neil S., and Paula Truax. 1991. Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research. *Journal of Consulting and Clinical Psychology* 59:12–19.

Joachim, Sabine, Matthew R. Sanders, and Karen M. T. Turner. 2010. Reducing Preschoolers' Disruptive Behavior in Public with a Brief Parent Discussion Group. *Child Psychiatry and Human Development* 41:47–60.

Karpenko, Veronika, Julie S. Owens, Nicole M. Evangelista, and Cassandra Dodds. 2009. Clinically Significant Symptom Change in Children with Attention-Deficit/Hyperactivity Disorder: Does It Correspond with Reliable Improvement in Functioning? *Journal of Clinical Psychology* 65:76–93.

Kazdin, Alan E. 1999. The Meanings and Measurement of Clinical Significance. *Journal of Consulting and Clinical Psychology* 67:332–39.

Kazdin, Alan. E. 2008. Evidence-Based Treatments and Delivery of Psychological Services: Shifting our Emphases to Increase Impact. *Psychological Services* 5:201–15.

Kendall, Philip C., Abbe Marrs-Garcia, Sanjay R. Nath, and Radley C. Sheldrick. 1999 Normative Comparisons for the Evaluation of Clinical Significance. *Journal of Consulting and Clinical Psychology* 67:285–99.

Kling, Åsa, Martin Forster, Knut Sundell, and Lennart Melin. 2010. A Randomized Controlled Effectiveness Trial of Parent Management Training with Various Degrees of Therapist Support. *Behaviour Therapy* 41:530–42.

La Greca, Annette M. (2005). Editorial. *Journal of Consulting and Clinical Psychology*, 73:3–5.

Marrs-Garcia, Abbe. 2010. What Does "Work" Mean? Reopening the Debate About Clinical Significance. *Clinical Psychology Science and Practice* 17:48–51.

McGlinchey, Joseph B., David C. Atkins, and Neil S. Jacobson. 2002. Clinical Significance Methods: Which One to Use and How Useful Are They? *Behavior Therapy* 33:529–50.

Morawska, Alina, and Matthew. R. Sanders. 2006. Self-Administered Behavioural Family Intervention for Parents of Toddlers: Effectiveness and Dissemination. *Behaviour Research and Therapy* 44:1839–48.

Morawska, Alina, Divna Haslam, Danielle Milne, Matthew R. Sanders. 2011. Evaluation of a Brief Parenting Discussion Group for Parents of Young Children. *Journal of Developmental and Behavioral Pediatrics* 32:136–45.

Nixon, Reginald D. V., Lynne Sweeney, Deborah B. Erickson, and Stephen W. Touyz. 2003. Parent-Child Interaction Therapy: A Comparison of Standard and Abbreviated Treatments for Oppositional Defiant Preschoolers. *Journal of Consulting and Clinical Psychology* 71:251–60.

Ogles, Benjamin M., Kirk M. Lunnen, and Kyle Bonesteel. 2001. Clinical Significance: History, Application, and Current Practice. *Clinical Psychology Review* 21:421–46.

Plake, Barbara S., James C. Impara, and Robert A. Spies. 2003. *The Fifteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Reid, M. Jamila, Carolyn Webster-Stratton, and Mary Hammond. 2007. Enhancing a Classroom Social Competence and Problem-Solving Curriculum by Offering Parent Training to Families of Moderate- to High-Risk Elementary School Children. *Journal of Clinical Child and Adolescent Psychology* 36:605–20.

Reyno, Sandra M., and Patrick J. McGrath. 2006. Predictors of Parent Training Efficacy for Child Externalizing Behavior Problems – a Meta-Analytic Review. *Journal of Child Psychology and Psychiatry* 47:99–111.

Sanders, Matthew R. 1999. Triple P-Positive Parenting Program: Towards an Empirically Validated Multilevel Parenting and Family Support Strategy for the Prevention of Behavior and Emotional Problems in Children. *Clinical Child and Family Psychology Review* 2:71–90.

Sanders, Matthew R., Carol Markie-Dadds, Lucy A. Tully, and William Bor. 2000. The Triple P – Positive Parenting Program: A Comparison of Enhanced, Standard, and Self-Directed Behavioral Family Intervention for Parents of Children with Early Onset Conduct Problems. *Journal of Consulting and Clinical Psychology* 68:624–40.

Silverman, Wendy K., and Stephen P. Hinshaw. 2008. The Second Special Issue on Evidence-Based Psychosocial Treatments for Children and Adolescents: A 10-Year Update. *Journal of Clinical Child and Adolescent Psychology* 37:1–7.

Tingey, Richard C., Michael L. Lambert, Gary M. Burlingame, and Nathan B. Hansen. 1996. Assessing Clinical Significance: Proposed Extensions to the Method. *Psychotherapy Research* 6:109–23.

Webster-Stratton, Carolyn. 2000. The Incredible Years Training Series. *OJJDP Bulletin*, June, 1–24.

Wise, Edward A. 2004. Methods for Analyzing Psychotherapy Outcomes: A Review of Clinical Significance, Reliable Change and Recommendations for Future Directions. *Journal of Personality Assessment* 82:50–59.

Martin Forster
mrmartinforster@gmail.com

Åsa Kling
asa.kling@psyk.uu.se

Knut Sundell
knut.sundell@socialstyrelsen.se