J Child Fam Stud (2012) 21:1041–1049
DOI 10.1007/s10826-012-9566-4

ORIGINAL PAPER

Utilizing Benchmarking to Study the Effectiveness of Parent–Child Interaction Therapy Implemented in a Community Setting

Shannon Self-Brown · Jessica R. Valente · Robert C. Wild · Daniel J. Whitaker · Rachel Galanter · Shannon Dorsey · Jenelle Stanley

Published online: 25 February 2012

Abstract Benchmarking is a program evaluation approach that can be used to study whether the outcomes of parents/children who participate in an evidence-based program in the community approximate the outcomes found in randomized trials. This paper presents a case illustration using benchmarking methodology to examine a community implementation of Parent–Child Interaction Therapy (CI-PCIT) utilized as a child maltreatment prevention effort. Data were collected from 83 parent–child dyads. Change scores were compared to treatment and control effect sizes aggregated from the PCIT literature. Pre-post results indicated significant positive changes in child behavior for CI-PCIT completers. Benchmarking analyses revealed that parents who completed CI-PCIT reported significantly greater positive child outcomes than the aggregate control group benchmark, and significantly less than observed in the treatment benchmark. A summary of decision points and implications for utilizing this methodology in the child maltreatment field are postulated.

Keywords Benchmarking · Child maltreatment · Parent–Child Interaction Therapy · Implementaiton

S. Self-Brown (✉) · J. R. Valente · R. C. Wild · D. J. Whitaker · J. Stanley
Institute of Public Health, Georgia State University, Atlanta, GA 30302, USA
e-mail: sselfbrown@gsu.edu

R. Galanter
Durham Exchange Clubs' Family Center, Durham, NC, USA

S. Dorsey
University of Washington, Seattle, WA, USA

Introduction

In recent years, there has been a significant push to implement evidence-based programs (EBPs) in agencies or state systems that work with parents at risk for, or who have perpetrated, child maltreatment (Bridge et al. 2008). Behavioral parent training programs (BPTs) are increasingly recognized by experts as an evidence-based approach that could benefit such families (Barth et al. 2005; Chaffin and Friedrich 2004; Whitaker et al. 2005). BPTs, originally developed as an intervention for child externalizing behavior problems, are based on the assumption that parenting skill deficits can be improved by providing parents with a repertoire of skills that help parents avoid negative interactions with their children, improve expectations of their children's behavior, and use behavioral management techniques that improve child behavior, and, consequently, reduce the likelihood of harsh discipline (Barth 2009). BPTs have four core components: (1) parenting problems are assessed, (2) parents are taught new skills, (3) parents practice applying the skills taught with their children, and (4) parents receive feedback about their use of the targeted skills (Barth et al. 2005).

Several BPTs, including Parent–Child Interaction Therapy (PCIT; Hembree-Kigin and McNeil 1995), The Incredible Years (Reid et al. 2003), and Triple-P (Sanders et al. 2003), are currently being promoted for child welfare populations (California Evidence Based Clearinghouse for Child Welfare 2009; National Center for Injury Prevention and Control 2004), as there is mounting evidence that these programs work effectively for reducing child maltreatment risk and recidivism with the child welfare system (Chaffin et al. 2004, 2011; Prinz et al. 2009; Webster-Stratton and Reid 2010). As BPTs are more widely implemented by "real-world" providers, under "real-world conditions" (e.g., high caseloads, lower levels of fidelity monitoring), it is important to understand whether child and family outcomes approximate the outcomes seen in randomized trials (Hunsley and Lee 2007). While it would be preferable to study family outcomes utilizing rigorous methodologies, the extensive fiscal and personnel resources required for such studies typically are not possible without grant funding. Thus, it is imperative that researchers consider other methodological approaches to evaluate community programs and implementation effectiveness, particularly methodologies that offer an opportunity to capitalize on state or federally-funded implementation efforts and cost-effectively evaluate outcomes.

Benchmarking has been recommended as a program evaluation approach for examining efficacious interventions implemented in community settings (Hunsley and Lee 2007; Minami et al. 2007). With this method, pre- and post-intervention data collected in community settings can be compared to intervention efficacy data from clinical trials, using four steps (Minami et al. 2007): (1) identify the intervention and population; (2) select/construct the "gold standard" benchmark, which can be single randomized controlled trials (RCTs) (Hunsley and Lee 2007; Weersing 2005), published meta-analyses (Neill 2003), or individually aggregated effect sizes (ESs) from multiple RCTs (Curtis et al. 2009; Minami et al. 2007; Weersing and Weisz 2002), with aggregated data being touted as the most accurate and reliable; (3) measure community outcomes that are similar to the benchmark and calculate the ESs; and (4) compare the ESs of the community and benchmark outcomes. Thus, benchmarking methods can provide a measure of program evaluation and intervention effectiveness with minimal costs and burden to community agencies.

While there is an emerging literature of benchmarking studies focused on evidence-based mental health interventions for youth and adults (e.g. Farrell et al. 2010; Minami 2008b; Oei and Boschen 2009; Weersing 2005; Weersing and Weisz 2002), there are no published studies to date using this methodology to benchmark a BPT implemented with parents at high risk for maltreatment. This paper provides an illustrative case example for the use of benchmarking methodology to evaluate a PCIT implementation targeting child abuse prevention with high risk parents in a "real world" community agency. This methodology was selected because a community agency implemented PCIT as part of their child abuse prevention services for at-risk parents and was interested in a rigorous evaluation of their program. Because no a priori evaluation was planned, the agency needed an evaluation method that could be completed through record review.

PCIT is a two-stage intervention for parent–child dyads, which includes a relationship enhancement phase (Child Directed Interaction [CDI]) and a discipline phase (Parent Directed Interaction [PDI]), to improve parenting skills and reduce child behavior problems (Hembree-Kigin and McNeil 1995). The efficacy of PCIT as a treatment for its original target problem—externalizing child behavior disorders—has been well established (Eyberg et al. 2005). Urquiza and McNeil (1996) first proposed a rationale for utilizing PCIT with maltreating parents, suggesting that physical abuse often occurs in a context of child discipline and a dysfunctional or detached parent–child relationship. Growing evidence suggests that an adapted version of PCIT, with an additional motivational component, can be an effective intervention for reducing recidivism for parents who have engaged in maltreatment, both when PCIT is delivered in a more strict, controlled setting by highly trained therapists (Chaffin et al. 2004) and by community therapists closely supervised by PCIT expert trainers (Chaffin et al. 2011).

For this study, the clinical records of families who participated in PCIT at a community family support agency were examined, and pre- and post-treatment data related to child externalizing behavior were collected [see Galanter et al. (in press) for a complete description of the program evaluation]. The support agency's mission is to provide child maltreatment prevention services and PCIT was implemented specifically to target parent skill-building among parents at risk. Child behavior problems were used as a proxy for reduction in child maltreatment risk for this study, as researchers have identified such problems as one of the four common co-occurring issues that is related to high-risk parenting and increases child maltreatment risk (Barth 2009). Based on the benchmarking literature (Curtis et al. 2009), it was hypothesized that the community implementation PCIT group (CI-PCIT) would be superior to an aggregate control benchmark (created from the control groups in six selected RCTs), with CI-PCIT families showing significantly greater improvement of parent reports of child behavior from pre- to post-treatment than families used as the control benchmark. Further, it was hypothesized that CI-PCIT families would have equivalent parent-reported child behavior outcomes to the treatment group benchmark derived from efficacy studies that used extensive procedures to maximize internal validity.

Method

Participants

Clinic records were reviewed for 83 families who enrolled in PCIT services at a community family support agency in a mid-size southeastern city between January 2007 and January 2009. The agency mission is to "Prevent child abuse through family support, counseling and education programs." The agency adopted PCIT for all families with children ages 2–10 who requested or were mandated to obtain in home parent support due to child abuse risk and poor parenting skills. The majority of parents were mothers (86.8%) with a mean age of 30 years (SD = 7.00). Parents were ethnically diverse, with 55.4% of parents self-identifying as Latina/o ($n = 46$), 37.3% as African American ($n = 31$), and 7.2% as Caucasian ($n = 6$). Ten percent of the parents were court mandated to receive services. Of the clinical files reviewed, 54 families (65%) completed PCIT services and 29 families did not complete services. To be identified as a PCIT completer, a parent has to exhibit mastery of the CDI skills and PDI skills according to standard PCIT criteria. Specifically, parents proceeded from the CDI phase to the PDI phase at the point they demonstrated 10 reflections, 10 labeled praises, and 10 behavior descriptions with their child, while also having 3 or less total questions, commands, or negative talk within 5 min. For a parent to complete the PDI phase of treatment, the parent had to demonstrate the use of 4 commands with their child, and at least 75% were required to be effective direct commands, following the correct sequence of consequences (positive or negative).

Independent t-tests revealed significant differences in ethnicity ($p < 0.01$), language ($p < 0.01$), and mandate for services ($p < 0.05$) between completers and non-completers. Completers tended to be Latino/a (64.8%), Spanish speaking (61.1%), and not mandated to receive services (92.6%) compared to non-completers (37.9% Latino/a, 37.9% Spanish speaking, and 31% mandated for services, respectively). No differences were revealed according to whether the parent was a mother or father. Overall, 68.5% of mothers completed services and 31.5% did not. In comparison, 60% of fathers completed and 40% did not.

Measures

Eyberg Child Behavior Inventory (ECBI; Eyberg and Pincus 1999)

The ECBI is a 36-item parent-report measure that assesses conduct behavior problems for children 2–16 years of age. The ECBI includes an Intensity scale, which measures the frequency of undesirable child behaviors on a 7-point Likert scale, and a Problem scale that measures whether or not the parent perceives each behavior as problematic. The reliability and validity of the ECBI is well-established (Eyberg and Pincus 1999). Specifically, the ECBI was standardized on 798 children with parents from six outpatient pediatric settings in the Southeast, and included children between ages 2 and 16 years. The composition of the sample corresponded closely to US Census data, with socioeconomic status evenly distributed. Thus, the sample had a broad range of ethnic and socioeconomic backgrounds, which allows for the applicability of the ECBI to children and adolescents in a wide variety of social and demographic contexts. This measure was completed by parents who participated in the PCIT program at pre- and post-treatment.

Study Design and Procedures

Benchmarking methodology was used to evaluate the effectiveness of the PCIT services on child behavior implemented by a community agency. Demographic and pre/post-test data on the ECBI completed as part of usual services were extracted from clinic files. Pre- to post-test differences were examined for intervention completers ($n = 54$), as well as intervention non-completers ($n = 11$) for whom post-assessment data was completed (there were 18 intervention non-completers for whom post-test data were not available). Next, change scores for completers on the ECBI measure were used to compute an aggregate ES for the study sample that was then compared to aggregate benchmark ESs computed by the study investigators from criterion-selected PCIT efficacy trials, as described below.

Criteria for Literature Search and Selection of Benchmarks

The following criteria were used for the literature review to identify appropriate PCIT studies to be included in the construction of aggregate benchmarks: (1) the ECBI was used as a primary outcome measure for the study; (2) the study was an RCT. Studies were excluded if the intervention included cultural adaptations for PCIT without a comparison condition of usual PCIT. Multiple literature searches in both PubMed and PsychInfo, using the aforementioned parameters, were completed.

In the first search, a meta-analysis of PCIT was identified (Thomas and Zimmer-Gembeck 2007), which included RCTs, nonrandomized trials, and single cohort studies with numerous outcome measures. The eight studies defined in this paper as an RCT included in the meta-analysis that utilized the ECBI were considered as possible benchmarks. Three of these RCTs were eliminated from consideration. Two were eliminated because they were follow-up evaluations from earlier RCT studies (Nixon et al. 2003), and one (McNeil et al. 1999) because when the research team reviewed the original paper, it was not clear that a randomized control design was employed (i.e. the methods stated participants were divided into two groups depending on therapist availability and there was no indication that randomization occurred in the study methods).

A second search focused on RCTs of PCIT published after 2004, as the Thomas and Zimmer-Gembeck (2007) meta-analysis only included published papers dated 2004 or earlier. Additional studies were identified; however, all but one study (McCabe and Yeh 2009) were eliminated because they involved adaptations to PCIT delivery and study participants were dissimilar from those in the current study [e.g. one study included children on the autism-spectrum (Solomon et al. 2008), another focused on Chinese families (Leung et al. 2009) and another on Puerto Rican families (Matos et al. 2009)].

Selected Studies for Benchmark Analyses

In sum, six total RCTs were selected (see Table 1). All PCIT treatment groups were included in the treatment benchmark, with the exception of the adapted PCIT treatment group included in the McCabe and Yeh (2009) study. More specifically, only the PCIT as usual group (and not the GANA group which involved a culturally adapted PCIT) was used from this study, which was deemed appropriate as the CI-PCIT sample included a high percentage of Latino families, who were primarily Spanish speakers receiving PCIT with no cultural adaptations. An additional note about McCabe and Yeh (2009) is that no pre-treatment data was reported, and, thus, the ES was determined using the published Cohen's d.

The control group benchmark included all randomized control groups in the selected studies (wait-list or treatment as usual), but did not include the social validation groups in the Nixon (2001) and Nixon et al. (2003) studies. It should be noted that none of the selected RCTs focused exclusively on the application of PCIT with maltreating families. The two RCTs (Chaffin et al. 2004, 2011) to date examining this intervention with maltreating parents did not include the ECBI as an outcome measure, and, thus, were excluded from consideration.

Table 1 Treatment and control raw score means, SD, and sample sizes for current study and benchmarks

Study	*N*	ECBI Intensity (raw scores)			ECBI Problem (raw scores)		
		M_{Pre}	M_{Post}	SD	M_{Pre}	M_{Post}	SD
Current study	54	134.3	94.08	33.65	18.07	8.02	9.10
Treatment benchmarks							
Eyberg et al. (1995)	10	159.5	117.5	16.6	20.7	6.6	4.8
Brestan et al. (1997)							
Mother	16	173	133	29.5	23	11	5.8
Father	9	169	137	24.1	22	14	3.3
Schuhmann et al. (1998)							
Mother	22	170.3	117.6	26.4	21.9	10.9	6.5
Father	12	159.6	126.8	25.2	20.5	10.2	5
Nixon (2001)	17	166.58	125.24	18.93	–	–	–
Nixon et al. (2003)							
Mother	17	166.59	125.24	18.93	–	–	–
Father	17	148.33	124.0	24.54	–	–	–
McCabe and Yeh (2009)	19	–	95.44	–	–	11.06	–
Control benchmarks							
Eyberg et al. (1995)	6	170.7	177.2	40.3	23	21.5	10.3
Brestan et al. (1997)							
Mother	13	176	170	30.2	24	24	5.4
Father	7	181	185	41.2	25	24	10.2
Schuhmann et al. (1998)							
Mother	20	172.9	169.7	25.8	21.2	22.1	6.1
Father	10	167.7	160.9	36.5	24.6	17.4	5.7
Nixon (2001)							
WL	17	173.82	148.35	22.72	–	–	–
Nixon et al. (2003)							
WL Mother	17	173.82	148.35	22.72	–	–	–
WL Father	17	147.47	134.13	25.97	–	–	–
McCabe and Yeh (2009)	18	–	118.5	–	–	15.8	–

Nixon (2001) and Nixon et al. (2003) did not measure ECBI Problem subscale

PCIT Training and Program Fidelity in the CI-PCIT Group

CI-PCIT therapists were trained by a primary PCIT trainer with 7 years of training experience, and a co-trainer with 3 years of experience. Training was delivered over a period of 10 months in 2007 and 2008, which included didactic instruction, modeling, and behavioral rehearsal with feedback from trainers; the initial training was 2-days, with two follow-up, one-day booster training sessions. The first ten sessions conducted by therapists post-workshop were monitored, either live or recorded, by the trainers to ensure competence and treatment integrity in service delivery. Following training completion in 2008, the locally-trained PCIT supervisor at the agency continued to conduct two fidelity monitoring sessions a year for each therapist. Further, over 20% of PCIT sessions had a reliability coder for CDI and PDI client skills, to ensure fidelity to the PCIT model during service delivery with families. Fidelity among therapists was reported to be very high by both the trainers, and the local PCIT supervisor.

PCIT was implemented as described by Hembree-Kigin and McNeil (1995) with two modifications for the targeted population (i.e., parents at risk for maltreatment). First, as indicated in work by Chaffin et al. 2004, 2011), the initial session included a focus on engagement, using motivational interviewing components. Second, consistent with the recommendations to reduce logistical barriers and improve family engagement in parenting services (Briss et al. 2000), the PCIT services were provided in the home setting, with adaptations for assessment and delivery as described in Ware (2008), Masse and McNeil (2008), and Galanter et al. (in press).

Data Analyses

Benchmarking methodology was used to evaluate the effectiveness of CI-PCIT services on child behavior. First, paired *t*-tests were conducted to measure pre-post changes on both the ECBI Intensity and Problem scales. Second, using the studies identified above, aggregate PCIT treatment and control benchmarks were created. Lastly, the aggregate ES of the ECBI measures within the CI-PCIT group was compared to the aggregate ECBI ESs for both the PCIT treatment and control group benchmarks (details below).

Results

Pre- to Post-Change ECBI in CI-PCIT Group

No differences emerged between baseline ECBI scores for intervention completers and noncompleters. For the ECBI Intensity scores, a significant decrease from pre- to post-treatment occurred for both completers ($t[53] = 7.56$, $p < 0.0001$) and non-completers ($t[10] = 2.98$, $p < 0.01$). Similar significant decreases were seen for the ECBI Problem scores for completers ($t[53] = 6.92$, $p < 0.0$) and non-completers ($t[10] = 2.27, p < 0.05$). Please see Table 1 for means and standard deviations.

Benchmarking Analyses

In accordance with the benchmarking calculations established by Hunsley and Lee (2007), and as described in other benchmarking studies in the mental health field (Curtis et al. 2009; Minami et al. 2007) the effect size for both subscales of the ECBI were calculated for the current study as well as for each of the chosen benchmark studies. Due to the small sample sizes for the benchmarked studies, the ES calculation included a sample size correction (Minami et al. 2008a).

The ESs for the ECBI subscales were then aggregated to produce a single ES for the CI-PCIT study. The same method was applied to the selected RCTs with each ECBI subscale ES combined into a single respective ES for the treatment and control groups across studies.

To test whether the CI-PCIT aggregate ES is clinically superior or equivalent to the control benchmark aggregate ES, a non-central *t* test was used with critical *t* values set at the 95th percentile. The difference between two ESs deemed clinically meaningful was set to be $\Delta = 0.2$ (Minami et al. 2008a, b).

All calculations and analyses were conducted using standard spreadsheet software and the statistical computing package, R, version 2.10.1 (2009). Table 1 displays the means, standard deviations, and sample sizes of CI-PCIT and the selected RCT benchmarks. Table 2 presents the aggregate ESs for the benchmarks as well as the total aggregate treatment and control ESs. When comparing the current community sample to the control benchmark, the community sample (ES = 1.13) was found to be statistically superior to the aggregate control benchmark (ES = 0.53, $p < 0.0001$). However, when compared to the treatment benchmark (ES = 1.71), the current sample (ES = 1.13) was found to be not equivalent and statistically inferior to the aggregate treatment benchmark ($p < 0.0001$).

Discussion

The primary purpose of this paper was to provide a case illustration, using a PCIT implementation by a community agency, of benchmarking methodology to examine program effectiveness in child maltreatment-focused

Table 2 Group and aggregate effect sizes for current study, treatment benchmarks, and control benchmarks

Benchmark studies	Treatment aggregate effect sizes	Control (WL/TAU) aggregate effect sizes
Eyberg et al. (1995)	2.48	0.13
Brestan et al. (1997)	1.56	0.09
Schuhmann et al. (1998)	1.66	0.32
Nixon (2001)	2.08	1.07
Nixon et al. (2003)	1.36	0.78
McCabe and Yeh (2009)	1.96	1.70
All studies combined	1.71 (CVs = 1.89–2.35)	0.53 (CV = 1.01)
Current study	1.13	

PCIT Parent–child interaction therapy, *WL* waitlist, *TAU* treatment as usual, all reported benchmarks used ECBI, either the Intensity subscale, the Problem subscale, or both

implementation research. Hypotheses were partially supported. The CI-PCIT group did emerge as superior to the control group benchmark, which included waitlist control and treatment as usual families from the efficacy trials. However, the community PCIT was inferior to, or not equivalent to, the gold-standard PCIT benchmark.

Although, this is not what was hypothesized, this finding is understandable based on previous implementation research, ESs tend to diminish when studies move from clinic-based efficacy trials to effectiveness trials in real world implementation (e.g., Curtis et al. 2004). Several factors likely contributed to the drop in ES observed in this community-based implementation. First, similar to other studies focusing on families with child maltreatment risk (e.g., Chaffin et al. 2004, 2011), PCIT was implemented in this community agency primarily as an intervention for parenting deficits rather than strictly for child behavior problems. This is in contrast to the vast literature on PCIT and the studies used for the benchmarking in this case example, which focuses on improving the behavior of children with significant externalizing problems; thus, the ECBI scores are much higher at baseline compared to the current sample. Unfortunately, in the few studies that have examined PCIT with maltreating families, the ECBI was not included as an outcome measure (at least in the published manuscript). Second, the population served by the community agency of study was a very diverse and high risk population (impoverished, ethnic minority, limited English); quite different from many of the families included in the clinical trials where strict inclusion/exclusion criteria are delineated (with the exception of McCabe and Yeh 2009). Third, there were modifications made to the treatment delivery setting. PCIT was delivered in the home rather than clinical setting, which could impact the overall efficacy. While there is emerging evidence that PCIT delivered in the home can lead to significant improvements in parent behavior and parent–child interaction (Ware et al. 2008; Galanter et al. in press), there are no randomized trials comparing family outcomes from these two delivery methods. Lastly, data in this study were collected during the initial implementation of PCIT in the agency, and prior studies have indicated that families served by newly trained professionals often fare less well than those served by experienced trainees and experts (Chaffin et al. 2004).

The findings from this case illustration suggest that benchmarking can be a useful tool for understanding program effects in the absence of rigorous evaluation method, which can be difficult and costly to employ in community settings. However, there is much work needed to improve the applicability of benchmarking for the child maltreatment field. More guidance and research are necessary in the following areas, which correspond to the 4 steps of benchmarking, outlined by Minami et al. (2007):

1. *Issue related to identifying intervention and population comparisons.* Currently, there is little guidance for how to apply benchmarking methods to study well-established interventions applied to a different presenting problem. The child maltreatment field is applying BPTs, which have been successful in treating child externalizing behavior problems, to families at risk for child maltreatment and reducing maladaptive parenting strategies. It is unclear whether using intervention studies that focus on a related, but different presenting problem (i.e. externalizing disorders vs. child maltreatment risk) is appropriate or relevant. A related concern is the effectiveness of benchmarking methodology when applied to populations (e.g., cultural groups) that were not included in existing efficacy trials. If gold standard benchmarks are determined to be appropriate only when the community agencies implement an intervention for the exact presenting problem and with similar populations for which there are efficacy trials available, then this methodological approach may be too restrictive for broad use in the child maltreatment field. Clearly more research and discussion is necessary for establishing such decision points.

Another issue related to the selection of an intervention benchmark is the measurement of fidelity. Without data indicating that therapists delivering both the intervention in efficacy trials and community settings are properly following program guidelines and procedures to ensure program quality, benchmarking comparisons have less meaning, especially when the implementation group is inferior to the treatment benchmark group (Carroll et al. 2007; Hermann et al. 2006; McLeod et al. 2009). A significant limitation in this study was that the fidelity data collected by the local PCIT supervisor was not quantifiable or criterion-based in a way that could be reported. On a related note, it is important to consider how adaptations/modifications should be measured when monitoring fidelity and using benchmarking methodology. There were modifications made in this implementation, however, the major modification was related to the delivery setting (home vs. clinic), and not to assessment, content or skill-based adaptations. More work is warranted to provide answers for how such issues should be considered within benchmarking and more broadly in effectiveness research.

2. *Issues related to selecting appropriate studies for the "gold standard" benchmark.* With regard to the existing literature, the current researchers followed the recommendations from Hunsley and Lee (2007) to create benchmarks from a collection of individual RCTs. Another option would have been to compare the group ES from the CI-PCIT group to the PCIT meta-analysis ES (Thomas and Zimmer-Gembeck 2007). However, this meta-analysis included results from quasi-experimental studies, single case studies, and RCTs as well as measures outside the scope of the current study; thus, benchmarking against more selective RCTs seemed a more robust option. Other researchers have also expressed difficulty choosing appropriate benchmarks, and determining when it is most appropriate to use a select group of studies versus an existing meta-analysis (e.g. Curtis et al. 2009). This is even more of an issue when there are few efficacy trials available. As benchmarking methodology becomes more standard across the behavioral sciences, further criteria and guidelines should be established to assist researchers in selecting the best benchmarking criteria for their particular project goals.

3. *Issues related to the measures used to calculate ESs.* In order to use benchmarking methodology, it is imperative that providers in community settings are utilizing standardized measures of behavioral change at pre-test and post-test. It would be helpful for researchers to examine the decision-making process for newly implementing agencies regarding measurement and program evaluation. Such decisions are likely influenced by the cost and time required to implement standardized measures, as well as whether purveyors of EBPs require the utilization of such tools as part of a certification and quality assurance process.

With regard to PCIT, there are generally two measures that are required during the delivery of the intervention, the ECBI, which measures child behavioral change, and the DPICS, which measures parent behavioral change and parent–child interaction. The current case example was limited to the ECBI for benchmarking analyses, which may have impacted the results of this study. The DPICS data was also collected by the community agency therapists in the implementation sample (for more information, see Galanter et al. in press). In published PCIT research, DPICS coding is almost always conducted by blinded coders, and thus, the comparison to coding completed by the treating clinicians could be biased. Reliability of coding done by therapist versus blind researchers is not yet established. In an ideal situation, where a PCIT benchmarking evaluation could be planned a priori, implementing agencies could be required to send random videotaped DPICS observations to expert blind coders. Having a valid DPICS measure scored according to research criteria would vastly expand the research studies that could have been included in the aggregate benchmark scores. In general, more research is needed to help determine if and when it is appropriate to benchmark on measures that are scored or utilized somewhat differently in the community setting as compared to efficacy studies.

4. *Issues with comparing and interpreting the ESs between implementation group outcomes and benchmark outcomes.* To date, it is unclear how generalizable findings are from benchmark studies, especially considering that such findings are vulnerable to all the various threats to internal validity produced by a single group design. Sample size and attrition issues can further inhibit the overall confidence with such findings. Minami et al. (2007) indicated that benchmarking must be interpreted cautiously when sample sizes are not sufficient. Small sample size can significantly affect the computed critical values used for comparing the community sample ES to the treatment or control benchmarks, with small sample sizes of less than 100 producing higher critical values that must be exceeded to claim clinical equivalence. However, with community based research, it could take years for small programs interested in evaluating and potentially benchmarking their program to serve 100 families. Using this community implementation as an example, it took two full years to enroll 83 families, of which 54 completed. While it is possible to correct ES calculations for small sample sizes, future research should identify the lower limit threshold that is necessary to conduct benchmarking studies.

Another issue with comparing community implementations to existing efficacy studies is that clinical trial outcomes are often calculated based only on those who completed the treatment protocol (completer samples), without intent-to treat information for all participants who were initially randomized into conditions. This can greatly increase the ES found for treatment groups, making it less likely to find equivalency between clinical trial treatment benchmarks and implementation samples. Thus, it is imperative that efficacy research provide attrition and descriptive data of the intervention and comparison sample groups, which could increase the external validity of clinic-based data (Stewart and Chambless 2009). Another helpful solution is that those who are benchmarking their own programs against existing trials should present data comparing their completers and dropouts at baseline to examine whether the two groups were different. This would provide some information on the extent of the problem of attrition in the sample being benchmarked. For the current study, no significant differences were found between completers and noncompleters on the ECBI at pre-treatment.

Conclusions

One of the most important questions in the child welfare and mental health fields is whether research-based treatments can be delivered in community settings and achieve or approach the outcomes seen in randomized trials (e.g., Hunsley and Lee 2007). The current research provides an illustrative case example for how to apply a benchmarking evaluation approach to examine the effects of a child maltreatment intervention implementation. Results from this study helped further inform the participating community agency about how their program is faring in comparison to PCIT programs implemented more strictly for externalizing problems, with high internal validity and less external validity. Reportedly, the outcomes of this study assisted the program director in making further programming and measurement decisions based on empirical data. As seen in this study, benchmarking methodology can be readily utilized by community practitioners to allow greater understanding of program effectiveness and transportability when control groups are not possible; however, additional research is necessary to further develop the decisions and criteria for the creation of field or measure-specific benchmarks. Until then, decisions about whether to use this approach to examine child maltreatment prevention programs should be made on case-by-case basis.

References

Barth, R. (2009). Preventing child abuse and neglect with parent training: Evidence and opportunities. *The Future of Children, 19*, 95–118.

Barth, R. P., Landsverk, J., Chamberlain, P., Reid, J. B., Rolls, J. A., Hurlburt, M. S., et al. (2005). Parent-training programs in child welfare services: Planning for a more evidence-based approach to serving biological parents. *Research on Social Work Practice, 15*, 353.

Brestan, E. V., Eyberg, S. M., Boggs, S. R., & Algina, J. (1997). Parent–Child Interaction Therapy: Parents' perceptions of untreated siblings. *Child & Family Behavior Therapy, 19*, 13–28.

Bridge, T., Massie, E., & Mills, C. (2008). Prioritizing cultural competence in the implementation of an evidence-based practice model. *Children and Youth Services Review, 30*, 1111–1118.

Briss, P. A., Zaza, S., Pappaioanou, M., Fielding, J., Wright-De Aguero, L., Truman, B. I., et al. (2000). Developing an evidence-based guide to community preventive services—methods. The task force on community preventive services. *American Journal of Preventive Medicine, 18*, 35–43.

California evidence-based clearinghouse for child welfare. (2009). *Usage Guide for the CEBC*. Retrieved from http://www.cachildwelfareclearinghouse.org.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*, 40.

Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. (2004). *Using evidence-based parenting programs to advance CDC efforts in child maltreatment prevention: Research brief.* Atlanta, GA: Author.

Chaffin, M., & Friedrich, B. (2004). Evidence-based treatments in child abuse and neglect. *Children and Youth Services Review, 26*, 1097–1113.

Chaffin, M., Funderburk, B., Bard, D., Valle, L. A., & Gurwitch, R. (2011). A combined motivation and Parent–Child Interaction Therapy package reduces child welfare recidivism in a randomized dismantling field trial. *Journal of Consulting and Clinical Psychology, 79*, 84–95.

Chaffin, M., Silovsky, J. F., Funderburk, B., Valle, L. A., Brestan, E. V., Balachova, T., et al. (2004). Parent–Child Interaction Therapy with physically abusive parents: Efficacy for reducing future abuse reports. *Journal of Consulting and Clinical Psychology, 72*, 500–510.

Curtis, N. M., Ronan, K. R., & Borduin, C. M. (2004). Multisystemic treatment: A meta-analysis of outcome studies. *Journal of Family Psychology, 3*, 411–419.

Curtis, N. M., Ronan, K. R., Heiblum, N., & Crellin, K. (2009). Dissemination and effectiveness of multisystemic treatment in New Zealand: A benchmarking study. *Journal of Family Psychology, 23*, 119–129.

Eyberg, S. M., Boggs, S. R., & 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., Nelson, M. M., Duke, M., & Boggs, S. R. (2005). Manual for the dyadic parent–child interaction coding system, 3rd edn. Gainesville: University of Florida. Available from http://www.PCIT.org.

Eyberg, S. M., & Pincus, D. (1999). *Eyberg Child Behavior Inventory Sutter Student Behavior Inventory—revised professional manual*. Odessa, FL: Psychological Assessment Resources.

Farrell, L. J., Schlup, B., & Boschen, M. J. (2010). Cognitive-behavioral treatment of childhood obsessive-compulsive disorder in community-based clinical practice: Clinical significance and benchmarking against efficacy. *Behaviour Research and Therapy, 48*, 409–417.

Galanter, R., Self-Brown, S., Valente, J., Dorsey, S., Whitaker, D., Bertuglia-Haley, M., & Prieto, M. (in press). Effectiveness of Parent Child-Interaction Therapy delivered to at-risk families in the home setting. *Child and Family Behavior Therapy.*

Hembree-Kigin, T., & McNeil, C. (1995). *Parent–Child Interaction Therapy*. New York: Plenum.

Hermann, R. C., Mattke, S., Somekh, D., Silfverhielm, H., Goldner, E., Glover, G., et al. (2006). Quality indicators for international benchmarking of mental health care. *International Journal for Quality in Health Care, 18*, 31–38.

Hunsley, J., & Lee, C. M. (2007). Research-informed benchmarks for psychological treatments: Efficacy studies, effectiveness studies, and beyond. *Professional Psychology: Research and Practice, 38*, 21–33.

Leung, C., Tsang, S., Heung, K., & Yiu, I. (2009). Effectiveness of Parent–Child Interaction Therapy (PCIT) among Chinese families. *Research on Social Work Practice, 19*, 304–313.

Masse, J. J., & McNeil, C. B. (2008). In-home Parent–Child Interaction Therapy: Clinical considerations. *Child and Family Behavior Therapy, 30*, 127–135.

Matos, M., Bauermeister, J. J., & Bernal, G. (2009). Parent–Child Interaction Therapy for Puerto Rican preschool children with ADHD and behavior problems: A pilot efficacy study. *Family Process, 48*, 232–252.

McCabe, K., & Yeh, M. (2009). Parent–Child Interaction Therapy for Mexican Americans: A randomized clinical trial. *Journal of Clinical Child & Adolescent Psychology, 38*, 753–759.

McLeod, B. D., Southam-Gerow, M. A., & Weisz, J. R. (2009). Conceptual and methodological issues in treatment integrity measurement. *Social Psychology Review, 38*, 541–546.

McNeil, C. B., Capage, L. C., Bahl, A., & Blanc, H. (1999). Importance of early intervention for disruptive behavior problems: Comparison of treatment and waitlist-control groups. *Early Education and Development, 10*, 445–454.

Minami, T., Serlin, R. C., Wampold, B. E., Kircher, J. C., & Brown, G. S. (2008a). Using clinical trials to benchmark effects produced in clinical practice. *Quality & Quantity, 42*, 513–525.

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., & Brown, G. S. (2008b). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. *Journal of Consulting and Clinical Psychology, 76*, 116–124.

Minami, T., Wampold, B. E., Serlin, R. C., Kircher, J. C., & Brown, G. S. (2007). Benchmarks for psychotherapy efficacy in adult major depression. *Journal of Consulting and Clinical Psychology, 75*, 232–243.

Neill, J. T. (2003). Reviewing and benchmarking adventure therapy outcomes: Applications of meta-analysis. *Journal of Experiential Education, 25*, 316–321.

Nixon, R. D. V. (2001). Changes in hyperactivity and temperament in behaviourally disturbed preschoolers after Parent–Child Interaction Therapy (PCIT). *Behaviour Change, 18*, 168–176.

Nixon, R. D. V., Sweeney, L., Erickson, D. B., & Touyz, S. W. (2003). Parent–Child Interaction Therapy: A comparison of standard and abbreviated treatments for oppositional defiant preschoolers. *Journal of Consulting and Clinical Psychology, 71*, 251–260.

Oei, T. P. S., & Boschen, M. J. (2009). Clinical effectiveness of a cognitive behavioral group treatment program for anxiety disorders: A benchmarking study. *Journal of Anxiety Disorders, 23*, 950–957.

Prinz, R. J., Sanders, M. R., Shapiro, C. J., Whitaker, D. J., & Lutzker, J. R. (2009). Population- based prevention of child maltreatment: The US Triple P System population trial. *Prevention Science, 10*, 1–12.

Reid, M. J., Webster-Stratton, C., & Hammond, M. (2003). Follow-up of children who received the Incredible Years intervention for oppositional-defiant disorder: Maintenance and prediction of 2-year outcome. *Behavior Therapy, 34*, 471–491.

Sanders, M. R., Markie-Dadds, C., & Turner, K. M. T. (2003). *Theoretical, scientific and clinical foundations of the Triple P-Positive Parenting Program: A population approach to the promotion of parenting competence*. St. Lucia, Australia: Parenting and Family Support Centre.

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.

Solomon, M., Ono, M., Timmer, S. G., & Goodlin-Jones, B. (2008). The effectiveness of Parent–Child Interaction Therapy for families of children on the autism spectrum. *Journal of Autism Developmental Disorders, 38*, 1767–1776.

Stewart, R. E., & Chambless, D. L. (2009). Cognitive-behavioral therapy for adult anxiety disorders in clinical practice: A meta-analysis of effectiveness studies. *Journal of Consulting and Clinical Psychology, 77*, 595–606.

R Development Core Team (2009) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

Thomas, R., & Zimmer-Gembeck, M. J. (2007). Behavioral outcomes of Parent–Child Interaction Therapy and Triple P-Positive Parenting Program: A review and meta-analysis. *Journal of Abnormal Child Psychology, 35*, 475–495.

Urquiza, A. J., & McNeil, C. B. (1996). Parent–Child Interaction Therapy: An intensive dynamic intervention for physically abusive families. *Child Maltreatment, 1*(2), 133–144.

Ware, L. M., McNeil, C. B., Masse, J., & Stevens, S. (2008). Efficacy of in-home Parent–Child Interaction Therapy. *Child & Family Behavior Therapy, 30*, 99–126.

Webster-Stratton, C., & Reid, M. J. (2010). Adapting the Incredible Years, an evidence-based parenting program, for families involved in the child welfare system. *Journal of Children's Services, 5*, 25–42.

Weersing, V. R. (2005). Benchmarking the effectiveness of psychotherapy: Program evaluation as a component of evidence-based practice. *Journal of American Academy of Child Adolescent Psychiatry, 44*, 1058–1062.

Weersing, V. R., & Weisz, J. R. (2002). Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. *Journal of Consulting and Clinical Psychology, 70*, 299–310.

Whitaker, D. J., Lutzker, J. R., & Shelley, G. A. (2005). Child maltreatment prevention priorities at the Centers for Disease Control and Prevention. *Child Maltreatment, 10*, 245–259.