Published in final edited form as:
Adm Policy Ment Health. 2017 September ; 44(5): 626–641. doi:10.1007/s10488-016-0769-7.

Benchmarking Family Therapy for Adolescent Behavior Problems in Usual Care: Fidelity, Outcomes, and Therapist Performance Differences

Aaron Hogue,
The National Center on Addiction and Substance Abuse

Sarah Dauber, and
The National Center on Addiction and Substance Abuse

Craig E. Henderson
Department of Psychology, Sam Houston State University

Abstract

This study evaluated whether community therapists delivering family therapy for adolescent behavior problems in usual care achieved performance benchmarks established in controlled trials for treatment fidelity and outcomes, with particular focus on individual differences in therapist performance. The study contained N = 38 adolescents (50% male; mean age 15 years) whose self-reported race/ethnicity was Hispanic (74%), African American (11%), multiracial (11%), and other (4%). Clients were treated by 13 therapists in one community mental health clinic that delivered family therapy as the routine standard of care. Therapists provided self-report data on adherence to core family therapy techniques; these scores were inflation-adjusted based on concordance with observer reports. Results showed that community therapists surpassed the fidelity benchmark for core family therapy techniques established by research therapists during a controlled trial. Regarding change in client functioning at six-month follow-up, community therapists were equivalent to the benchmark for internalizing symptoms and superior for externalizing symptoms and delinquent acts. Community therapists also demonstrated a high degree of performance uniformity: Each one approximated the fidelity benchmark, and only two produced relatively weak outcomes on any of the client change indicators. Caveats for interpreting therapist performance data, given the small sample size, are described. Recommendations are made for developing therapist-report fidelity measures and utilizing statistical process control methods to diagnose therapist differences and enhance quality assurance procedures.

Correspondence concerning this article should be addressed to Aaron Hogue, Director of Adolescent and Family Research, The National Center on Addiction and Substance Abuse, 633 Third Avenue, 19th floor, New York, NY, 10017. ahogue@centeronaddiction.org.

Dr. Hogue declares that he has no conflict of interest. Dr. Dauber declares that she has no conflict of interest. Dr. Henderson declares that h has no conflict of interest.

Ethical approval: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Ethical approval: This article does not contain any studies with animals performed by any of the authors.

Informed consent: Informed consent was obtained from all individual participants included in the study.

Keywords

benchmarking; therapist effects; family therapy; usual care; fidelity; outcomes; adolescent behavior problems

This study evaluated whether community therapists delivering family therapy for adolescent behavior problems in usual care reached performance benchmarks established by research-trained therapists implementing manualized family-based treatments. Benchmarks included both treatment fidelity, in the form of adherence to core family therapy techniques for this population; and treatment outcomes, in the form of six-month reductions in delinquent acts, externalizing symptoms, and internalizing symptoms. A main focus was examining individual differences among community therapists in fidelity and outcome performance using statistical process control, a method that has enormous utility for quality assurance purposes in routine care.

Benchmarking the Implementation of Evidence-Based Treatments in Usual Care

Benchmarking is a method for assessing whether therapists delivering evidence-based treatments (EBTs) in routine clinical settings can approximate performance standards set by research-funded clinicians in controlled trials (Spilka & Dobson, 2015). Benchmarking studies typically focus on critical areas such as client retention, model implementation, and clinical outcomes (Hunsley & Lee, 2007). The procedural steps in EBT benchmarking analyses are straightforward: (1) define the clinical problem, client population, and treatment model of interest; (2) identify (or calculate) “gold-standard” performance criteria from a relevant databased source; (3) measure therapist performance in an applied setting using methods comparable to those used to establish the benchmarks; (4) directly compare applied performance to databased benchmarks and explore reasons for observed discrepancies (Weersing, 2005).

One EBT ripe for benchmarking research on treatment fidelity and client outcomes in usual care is family therapy for adolescent behavior problems. Manualized family therapy (FT) models have produced an exemplary record of treatment effectiveness across the adolescent behavioral health spectrum and have reached the highest levels of empirical validation for disruptive behavior (Chorpita et al., 2011; Henggeler & Sheidow, 2012) and substance use (Hogue, Henderson, Ozechowski, & Robbins, 2014; Tanner-Smith, Wilson, & Lipsey, 2012). Studies have also consistently reported reductions in internalizing symptoms and gains in prosocial functioning (Hogue & Liddle, 2009). The large accumulation of controlled research on FT for adolescent behavior problems has facilitated well-powered meta-analytic reviews (e.g., Baldwin, Christian, Berkeljon, Shadish, & Bean, 2012; Tanner-Smith et al., 2012) that offer sturdy benchmarks for outcome success. As described below, the current study utilized research-derived benchmarks to gauge the performance of family therapists treating adolescents in routine practice conditions, that is, without the (presumed) benefits of extramural training and supervision in a specific manualized model.

Therapist Differences in Efficacy versus Implementation Research on EBTs

In addition to benchmarking analyses, this study examined therapist differences in observed FT fidelity and outcomes. In traditional controlled efficacy research, the primary goal is maximize therapist homogeneity in treatment delivery and outcomes in order to establish the potency of the model qua model: How effective is the treatment itself? In this context therapist performance differences are considered a nuisance factor—the model qua therapist—that obscures interpretation of model effects. This prompts efforts to mitigate therapist differences on two fronts. First, efficacy trials institute standardized fidelity procedures for selecting, training, and supervising therapists in order to generate performance similarity among multiple therapists within a given study condition (Elkin, 1999). Second, two complementary statistical procedures are employed to control for therapist differences in client outcomes, typically called “therapist effects” (see Crits-Christoph & Mintz, 1991). To correct for mean-level differences among therapists in outcomes (i.e., therapist main effects), a Therapist variable is modeled as a between-subjects fixed factor in study analyses (e.g., Project MATCH Research Group, 1998). To correct for inter-correlations among multiple clients treated by a single therapist (i.e., therapist nesting effects), which can lead to overestimation of group differences (Wampold & Serlin, 2000), mixed effects analyses are used that incorporate Therapist as a random factor (Zucker, 1990).

Treatment implementation research offers a very different context for conceptualizing and analyzing therapist differences. The aim of implementation science is to elucidate the conditions under which efficacious treatments can be delivered with fidelity by front-line therapists and sustained over time in community settings (McHugh & Barlow, 2010). In this context model efficacy is premise rather than goal, and the scientific focus shifts to understanding how multiple interrelated factors—client, therapist, provider, service system—interact to facilitate or inhibit effective model implementation. As a result, nuisance factors transform into main events. For example, therapist differences in model aptitude, training outcomes, fidelity success, and client outcomes frequently take center stage in EBT implementation studies (Beidas & Kendall, 2010).

Concerns about when and how therapist differences emerge during EBT delivery in everyday practice fall in the province of quality assurance (QA; Bond, Becker, & Drake, 2011). The counterpart to treatment integrity procedures in efficacy research, QA procedures are designed to ensure that EBTs adopted in routine care are implemented in accordance with the main principles and procedures of the given model (Schoenwald, 2011). QA procedures are tailored to fit the therapeutic content, administration requirements, and fidelity monitoring needs of the given EBT; as such, providers must be judicious in selecting effective yet resource-friendly procedures to sustain ongoing EBT delivery (Hogue, Ozechowski, Robbins, & Waldron, 2013). This study advances the literature on QA resources by illustrating the utility of process control benchmarking for routine QA purposes, especially for evaluation of therapist differences.

Study Background and Innovations

The current study draws on archived data from a randomized trial of usual care interventions for adolescent behavior problems (Hogue, Dauber, et al., 2014). The parent trial assigned 204 teens referred for conduct or substance use problems to either usual care family therapy (UC-FT, described in Methods section) or non-family treatment (UC-Other). At one-year follow-up across the full sample, adolescents showed significant declines in youth-reported externalizing and internalizing symptoms, caregiver-reported externalizing and internalizing symptoms, and delinquent acts. UC-FT produced greater reductions than UC-Other in youth-reported externalizing and internalizing symptoms; also, among substance-using youth, UC-FT had greater reductions than UC-Other in both delinquent acts and substance use.

The parent trial demonstrated that UC-FT was effective in treating multiple adolescent behavior problems and was also comparatively stronger than alternative treatment approaches in several domains. These results beg an intriguing question that is fundamental to QA goals for EBTs: Can community therapists providing FT in everyday care approximate the lofty standards of treatment integrity and outcomes established by manualized FTs in research settings? Absent a randomized trial directly comparing UC-FT to a manualized FT (for a similar example see Weisz et al., 2012), this question can be approached via cost-efficient benchmarking methods that offer a practice-relevant perspective on FT performance.

This study featured five innovations to grow the knowledge base on EBT delivery in usual care. First, the parent trial recruited from a network of school- and community-based referral sources rather than from existing clinic referral streams. This strategy yielded a sample of “unmet need” adolescents: teens with significant behavioral health impairments who are not involved in the treatment system (Ozechowski & Waldron, 2010). Understanding the clinical needs of adolescents with behavior problems who do not typically cross the treatment threshold is critically important for designing inclusive and responsive behavioral care (Institute of Medicine, 2006). Second, the measurement of UC-FT fidelity focused on therapist use of core FT elements rather than adherence to a standardized FT manual. Core EBT elements refer to discrete treatment techniques shared across multiple treatment models for a given disorder (Chorpita, Becker, & Daleiden, 2007; Garland, Hawley, Brookman-Frazee, & Hurlburt, 2008). As such, core EBT elements: are approach-specific (i.e., identified with a particular treatment orientation) but model-free (i.e., not inextricably bound to a single manual/version); can be selectively applied to cases presenting with comorbid disorders and other diagnostic complexities (Barth et al., 2014); and suitably represent non-manualized clinical practices favored in routine care (Garland, Bickman, & Chorpita, 2010).

Third, UC-FT fidelity to FT was assessed with a therapist self-report tool. Therapist-report measures of EBT fidelity have several methodological strengths that strike a desirable balance between rigor and relevance in practice settings: they are quick, inexpensive, and non-intrusive; they capture the unique viewpoint of the provider delivering the interventions; and they can be completed throughout treatment, which facilitates measurement of infrequent but clinically meaningful interventions (Carroll, Nich, & Rounsaville, 1998;

Weersing, Weisz, & Donenberg, 2002). Therapist-report measures can also enrich routine QA via feedback loops of several kinds, as exemplified in the current study: as a self-check by therapists to mark their own progress in treating cases; a supervision aid for trainers to monitor fidelity; and administrative data for reviewers to evaluate therapist- and agency-level performance (Schoenwald, Letourneau, & Halliday-Boykins, 2005). A previous study documented that the UC-FT therapists in the current sample were reliable in reporting on their own adherence to core FT techniques (Hogue, Dauber, Lichvar, Bobek, & Henderson, 2015; see *Fidelity Measure* in Method section).

Fourth, this study statistically corrected the fidelity scores reported by UC-FT clinicians in order to account for systematic overestimation (i.e., score inflation) in self-ratings of their own adherence to FT. That is, whereas our previous research found that UC-FT clinicians were *reliable* reporters of fidelity, they were not *accurate* ones. As is true in every study of therapist-report EBT adherence to date involving either research-hired (e.g., Carroll et al., 1998; Martino et al., 2009) or agency-hired (e.g., Brosan et al., 2008; Hurlburt et al., 2010) clinicians, our previous analyses of this study sample (Hogue, Dauber, Lichvar, et al., 2015) found that UC-FT therapists reported a significantly higher mean FT score than that reported by observational raters, over-reporting by four-tenths of a scale point on average. In other words, UC-FT therapists reliably documented their relative use of FT techniques in any given session—more versus less FT—but overstated the quantity of FT delivered. Self-report fidelity score inflation appears deeply rooted in benign reporter biases of several kinds, for example, perceived effort in delivering an intervention and/or a more inclusive framework for evaluating an intervention. To compensate for this essentially intractable reporting bias, the current study levied a sample-specific inflation-adjustment correction of four-tenths of a scale point (as fully described in *Plan of Analysis*), matching the known overestimation in self-report FT scores recorded by UC-FT.

Fifth, this study used statistical process control (SPC) analyses to articulate therapist differences in achieving fidelity and outcome benchmarks. SPC was developed in industrial psychology to monitor variability in a continuous production process (Deming, 1986) and is also an efficient and flexible approach for conducting benchmarking analyses during routine QA (Hogue et al., 2013). SPC employs probability sampling procedures (see Weersing & Weisz, 2002) in which continuous samples are taken from a process and plotted on a control chart containing upper and lower control limits based on either pre-specified criterion values or the distributive properties of the given sample. Plotted data points are then inspected to identify outliers and/or determine whether an "out of control" pattern emerges to signal a systematic change in the production process (Hoyer & Ellis, 1996). The current study complemented a conventional analysis of therapist similarity in fidelity and outcomes—calculation of an intraclass correlation coefficient (ICC) representing the ratio of between-therapist variance to total variance (Adelson & Owen, 2012)—with follow-up SPC analyses that plotted individual therapist performance against established FT benchmarks.

Current Study: Related Literature, Benchmark Sources, and Specific Aims

This study investigated the performance of community family therapists treating adolescents in usual care, comparing their FT fidelity scores and client outcomes to research-based benchmarks and exploring individual therapist differences. There has been only a handful of fidelity benchmarking studies for the FT approach, with some finding comparability (e.g., Hogue & Dauber, 2013) and others discrepancy (e.g., Henggeler, Pickrel, & Brondino, 1999) between target versus benchmark performances. One outcomes benchmarking study of a manualized FT (Curtis, Ronan, Heiblum, & Crellin, 2009) reported that community therapists implementing multisystemic therapy (MST) produced outcomes similar to research-derived MST benchmarks on several indicators, including juvenile offenses, out-of-home placement, and school/vocational attendance (Curtis, Ronan, Heiblum & Crellin, 2009). A more robust literature exists on therapist differences among FT models. In the fidelity domain, therapist differences have found in perceptions about model implementation difficulty (Schoenwald et al., 2005), time required among trainees to achieve benchmark fidelity scores (Schoenwald, Henggeler, Brondino, & Rowland, 2000), and fidelity score discrepancies between training cohorts in a dissemination study (Lofholm, Eichas, & Sundell, 2014). In the outcomes domain, therapist differences have been related to therapist-client ethnic match (Flicker et al., 2008), job satisfaction (Schoenwald, Chapman, Sheidow, & Carter, 2009), and perceptions about participatory decision-making (Schoenwald, Carter, Chapman, & Sheidow, 2008), to name a few.

For the fidelity benchmark we used core element FT adherence scores generated by Hogue, Dauber, Samuolis, and Liddle (2006) from an efficacy trial of manualized FT for adolescent behavior problems. The Hogue et al. study provides the only applicable benchmark data for the UC-FT sample, in that its observational FT adherence measure was directly translated into the therapist-report measure completed by UC-FT therapists (see *Fidelity Measure* below). Using benchmark data from a single study—known as point-by-point benchmarking—is acceptable under these conditions (Hunsley & Lee, 2007; Weersing, 2005). For the outcomes benchmark we used an averaged effect size reported by Baldwin and colleagues (2012) in their meta-analysis of controlled trials for adolescent behavior problems involving four manualized FT models (see *Benchmark Data Sources* below). This meta-analysis contains outcome benchmark data that are particularly apt for the heterogeneous UC-FT sample: Outcome variables are aggregates of conduct problems, substance use, and secondary outcomes such as externalizing symptoms, internalizing symptoms, and school performance.

The primary study aim was to test the ability of community therapists to reach fidelity and outcome benchmarks for the FT approach reported in controlled studies. This aim pertains to the feasibility and potency of EBTs when implemented in standard practice. A secondary aim was to illustrate the utility of statistical process control analysis as a resource-efficient QA method for tracking therapist differences in everyday care. Based on a previous study of FT fidelity achieved by an earlier cohort of therapists working at the UC-FT site (Hogue & Dauber, 2013; fully described in Method section), we expected that the sample pool of UC-FT therapists would approximate the FT fidelity benchmark but also demonstrate notable heterogeneity. Based on outcomes from the parent trial (Hogue, Dauber, et al., 2014), we expected that UC-FT therapists would collectively achieve the FT outcome benchmark but again demonstrate notable individual differences.

Method

The parent randomized trial from which these data were collected was conducted between 2006–2012 under approval by the governing Institutional Review Board.

Study Clients

The study sample (n = 38) contains all clients from the UC-FT condition of the parent trial for whom there was at least one therapist-report checklist. Study clients were adolescents (50% male; mean age 15.3 years [SD = 1.6]) and their primary caregivers. Self-reported race/ethnicity was Hispanic (74%), African American (11%), multiracial (11%), and other (4%). Households were headed by single parents (68%), two parents (22%), or grandparents (10%). A total of 60% of caregivers graduated high school, 56% were employed, 56% earned less than $15,000 per year, and 16% currently received public assistance. Adolescents were referred to the parent study from schools (76%) or other sources (24%); 19% were involved in the juvenile justice system at referral. At baseline study clients reported an average of 3.0 (SD = 3.4) delinquent acts and 3.6 (SD = 8.0) days of substance use in the prior month.

Rates of psychiatric diagnosis were assessed with the Mini International Neuropsychiatric Interview (Version 5.0; Sheehan et al., 1998), based on the fourth edition of Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR; American Psychiatric Association, 2000). Psychiatric diagnoses were given for meeting symptom thresholds based on either adolescent or caregiver report, with the following rates: Oppositional Defiant Disorder (ODD) = 91%, Attention-Deficit/Hyperactivity Disorder = 77%, Conduct Disorder (CD) = 50%, Mood Disorder or Dysthymia = 46%, Substance Use Disorder (SUD) = 27% (16% cannabis use, 16% alcohol), Generalized Anxiety Disorder = 14%, Posttraumatic Stress Disorder = 16%. A total of 83% of the sample was diagnosed with more than one disorder.

The study sample of 38 was derived from 104 participants randomized to the UC-FT condition in the parent trial; it represents 13 therapists who submitted 458 total self-report FT adherence checklists. There were no significant differences between the study sample versus the remaining pool of 66 UC-FT cases in the parent trial on any demographic or diagnostic variable.

Client Recruitment, Assessment, and Enrollment in the Parent Trial

Clients were part of a randomized trial designed to identify adolescents with untreated behavioral health problems, enroll them in available outpatient treatment services, and assess treatment effects. Research staff developed a referral network of high schools, family service agencies, and youth programs serving a large inner-city area. Staff contacted referred families by phone and offered them an opportunity to participate in home-based interviews to assess the reasons for study referral and discuss treatment enrollment. After completion of a baseline interview, adolescents who met diagnostic criteria for ODD, CD, or SUD and whose families were interested in receiving treatment were randomly assigned to UC-FT or an alternative study condition (UC-Other) comprised of five clinics that did not feature FT as a primary therapy approach. Of the 104 parent trial cases in the UC-FT condition, 77 (74%)

attended a clinical intake session; for this subgroup the average number of completed treatment sessions was 8.7 (SD = 9.8). Of the 77 UC-FT cases who completed a clinical intake session, 43 (56%) attended at least one treatment session post-intake; for this subgroup the average number of completed sessions was 14.2 (SD = 10.1). These rates are comparable to treatment engagement rates broadly reported for child mental health services (Garland et al., 2013).

Study Site and Therapists

The UC-FT treatment site was a community mental health center that accepted study cases as standard referrals. The site featured core family therapy techniques (Haley, 1987; Minuchin & Fishman, 1981) as the foundational approach for behavioral interventions with youth. At no time had the site imported a manualized FT model or contracted for extramural implementation support. No external training or financial support of any kind was provided to treat study cases, and therapists were not asked to alter their clinical practices in any way. All site therapists who volunteered to participate were accepted into the study; approximately 75% of the full-time staff and 30% of the part-time staff joined (N.B.: a majority of part-time staff were student trainees with insufficient time and caseload capacity to participate). The site generally prescribed weekly therapy sessions and offered in-house psychiatric support.

All site therapists received regular in-house training and supervision to promote family-based case conceptualization and use of signature treatment techniques of the FT approach; this included weekly meetings attended by all clinical staff for review and disposition of intakes, monthly educational seminars and trainings in various facets of the FT approach, and weekly individual supervision by licensed clinicians with varied expertise in a range of FT models. UC-FT therapists (N = 13, who treated 38 cases total) were licensed Marriage and Family Therapists, social workers with training in family therapy, or advanced trainees with family therapy experience. The 13 study therapists represent all but one of the 14 therapists in the parent trial (Hogue, Dauber, Henderson, et al., 2015); one of the original 14 provided no therapist-report fidelity data and two had incomplete outcome data. Participating therapists ranged from 28 to 59 years old; 78% were female and 89% Hispanic; and as a group averaged 3.1 years (SD = 4.3) postgraduate therapy experience. Demographic data were unavailable for 4 of the 13 therapists.

UC-FT Treatment Fidelity

Previous fidelity evaluations—Two previous studies of fidelity within the parent trial evaluated treatment adherence and differentiation for the UC-FT condition. One study using therapist-report data (Hogue, Dauber, & Henderson, 2014) found that, compared to UC-Other clinicians (who all practiced non-family approaches), UC-FT clinicians reported stronger allegiance and skill in FT techniques prior to treating study cases and greater utilization of FT techniques than techniques associated with cognitive-behavioral therapy (CBT), motivational interviewing (MI), or drug counseling (DC) while treating study cases. A follow-up observational fidelity study (Hogue, Dauber, Lichvar, et al., 2015) reported that UC-FT sessions evidenced greater use of FT techniques than MI/CBT techniques, even after controlling for therapist effects. Finally, at the end of each session study therapists were asked to document that session's format (Individual/Family versus Group) and participants (adolescent, caregiver, and/or other person); previous research (Hogue, Dauber, Henderson, & Liddle, 2014) indicates that therapists are highly reliable in documenting these structural features of treatment delivery. Within the UC-FT condition 100% of sessions were listed as Individual/Family; also, 75% of all sessions included the adolescent, 68% a caregiver, and 15% another person.

Previous benchmarking study—Hogue & Dauber (2013) used observational analyses to compare FT fidelity scores for archived sessions from the UC-FT site to benchmark scores from an efficacy trial of a manualized FT, multidimensional family therapy (MDFT; for a summary of MDFT evidence see Hogue, Henderson, et al., 2014). This exploratory study included 15 randomly selected sessions videotaped on site prior to the start of the parent trial by a previous cohort of therapists. Sessions were rated for adherence to core FT techniques using a well-validated observational fidelity tool (Hogue et al., 2006; the tool is described below in *Fidelity Measure*). SPC analyses (Deming, 1986) were used to plot within-sample variance in mean FT adherence scores for the UC-FT site against adherence data from the MDFT efficacy trial. The FT adherence tool contained a 7-point Likert-type rating scale: 1 = *Not at all*, 3 = *Somewhat*, 5 = *Considerably*, and 7 = *Extensively*. Scores for the archived UC-FT sessions (M = 3.4, SD = .51) clustered closely around the average score for MDFT (M = 3.5, SD = .60), and no score for any UC-FT session fell beyond two standard deviations of the benchmark MDFT mean. These analyses indicate that treatment delivered at the UC-FT site by a previous cohort of therapists adhered closely to gold-standard fidelity levels for signature FT techniques.

Fidelity Measure

The *Inventory of Therapy Techniques—Adolescent Behavior Problems (ITT-ABP)* is a 25-item QA tool designed to collect post-session therapist-report data on delivery of discrete treatment techniques associated with the FT, CBT, MI, and DC approaches. Inventory items were derived from validated observational fidelity scales for these respective EBTs using an instrument development process detailed in Hogue, Dauber, and Henderson (2014). The ITT-ABP assesses thoroughness/frequency with which each treatment technique was implemented in a just-completed session based on a 5-point Likert-type scale: 1 = *Not at all*, 2 = *A little bit*, 3 = *Moderately*, 4 = *Considerably*, 5 = *Extensively*. Construct validity was established via principal components analysis (PCA) on half of 822 inventories collected during the parent trial, followed by confirmatory factor analysis on the remaining half that yielded adequate fit indices: $\chi 2$ (272) = 388.01, $p < .001$; RMSEA = .03 (90% CI: .025–.039); CFI = .96; TLI = .96. This process identified three clinically coherent scales with strong internal consistency: FT scale (8 items: PCA item-factor loading range .73 – .46, Cronbach's α = .79), MI/CBT scale (8 items: PCA range .81 – .52, α = .87), and DC scale (9 items: PCA range .97 – .44, α = .90). The 8 core FT technique items are: Established definite theme/agenda at beginning of session; Discussed parental monitoring and family rules with the adolescent and/or caregiver; Worked individually with adolescent or caregiver to prepare for an in-session family interaction; Arranged, coached, and helped process a family interaction; Worked to enhance communication and attachment among family members; Shared information about normative adolescent development; Discussed core relational family themes that underlie everyday events (e.g., love, trust, respect, independence); Targeted intervention efforts at a family member participating in session.

The reliability of therapist-report ratings on the FT scale of the ITT-ABP was confirmed by observational coders utilizing the same 5-point rating scale (Hogue, Dauber, Lichvar, et al., 2015). Therapists and observers independently provided FT scale ratings on 157 sessions from both the UC-FT and UC-Other conditions of the parent trial. Therapist concordance with observers on averaged FT scale ratings was adequate (ICC = .66); moreover, UC-FT therapists and UC-Other therapists were comparably reliable in reporting on their use of FT techniques. One FT scale item (Established definite theme/agenda at beginning of session) was dropped from the current study due to its unacceptably low item-level interrater reliability (ICC = −.62). In terms of self-report accuracy, however, results were less promising: Compared to observers (M = 2.0; SD = .45), UC-FT therapists (M = 2.4; SD = .63) reported a significantly higher mean level of FT techniques. This overestimation bias prompted the use of inflation-adjustment procedures when analyzing the UC-FT fidelity data in the current study, described below in *Plan of Analysis*.

Outcome Measures

Measures used in the current study were administered at baseline and 6-month follow-up. The current study did not include a SUD outcome measure due to power limitations, as less than one-third of the parent trial reported substance use problems at baseline.

Externalizing and Internalizing symptoms—Adolescent reports of behavioral symptoms were assessed via the *Youth Self Report* (YSR). The YSR is supported by extensive evidence of reliability, validity, and clinical utility (Achenbach & Rescorla, 2001) and used with a wide range of adolescent samples. Total scores on the externalizing (oppositionality, aggression) and internalizing (depression, anxiety, somatization) summary scales were analyzed in this study.

Delinquent acts—Adolescent delinquency was assessed using the *National Youth Survey Self-Report Delinquency Scale* (SRD; Elliott, Huizinga, & Ageton, 1985), a well-validated instrument that has been used extensively with adolescent clinical samples. Adolescents reported on the number of times they engaged in various overt and covert delinquent acts.

Benchmark Data Sources

Fidelity benchmark—Benchmark FT fidelity scores were derived from the Hogue and colleagues (2006) observational analysis of FT adherence scores registered by MDFT during an efficacy trial that included 141 sessions from 63 families treated by five MDFT therapists. FT fidelity scores in that study were generated from observer ratings of the extensiveness (i.e., thoroughness and/or frequency) with which core FT techniques were implemented in each session; observers rated the same seven scale items that were subsequently included in the FT scale of the therapist-report ITT-ABP (described above in *Fidelity Measure*). Observer ratings for the MDFT sample used a 7-point scale with the following anchors (described above in *UC-FT Treatment Fidelity*): 1 = *Not at all*, 3 = *Somewhat*, 5 = *Considerably*, and 7 = *Extensively*. To harmonize benchmark scores with target scores

(Curran & Hussong, 2009), these MDFT ratings were transformed to the 5-point ITT-ABP scale used by UC-FT therapists: 1 = *Not at all*, 2 = *A little bit*, 3 = *Moderately*, 4 = *Considerably*, 5 = *Extensively*. Transformation of MDFT ratings from the 7-point to the 5-point scale proceeded as follows: scores of 1, 2, and 3 were retained; scores of 4 were changed to “3”; scores of 5 and 6 were changed to “4”; and scores of 7 were changed to “5”. This transformation resulted in a new score distribution that was highly similar to the distribution for the original 7-point scale. The transformed benchmark fidelity scores for observer-based ratings of core FT techniques in the MDFT sample were: M = 2.0, SD = 0.35.

Outcomes benchmark—The benchmark outcome score was derived from a meta-analysis completed by Baldwin and colleagues (2012) that analyzed 24 randomized studies testing one of four manualized FTs for adolescent problem behaviors: brief strategic family therapy, functional family therapy, MDFT, or MST. The benchmark outcome score is an effect size (ES) calculated as a standardized mean difference statistic corrected for small sample bias (Hedges & Olkin, 1985). The benchmark ES was aggregated across all client outcomes reported in each study, including primary (delinquency, substance use) and secondary (externalizing and internalizing symptoms, school attendance, etc.) outcome variables, using data collected during the first post-treatment assessment. The current study utilized the outcome ES calculated for manualized FT comparisons to usual care conditions across 11 separate studies: Cohen’s (1988) $d = .21$.

Plan of Analysis

Before analyzing the therapist-report FT adherence data provided by UC-FT clinicians, we adjusted these scores to correct for self-report inflation, as follows: We subtracted 0.4 from the mean FT scale score of each submitted ITT-ABP checklist, to compensate for sample-specific bias previously detected in the UC-FT ratings (as noted in the *Fidelity Measure* section). Specifically, an earlier study with the current sample (Hogue, Dauber, Lichvar, et al., 2015) found that UC-FT therapists reported a significantly higher mean level of FT techniques (M = 2.4; SD = .63) than did observational coders (M = 2.0; SD = .45) for the same set of sessions, thereby inflating their scores by an average of four-tenths of a scale point. This simple algebraic adjustment has the virtue of transparency, though it can be applied only to samples that have both therapist- and observer-report fidelity data on a common metric.

To conduct fidelity benchmark analyses we used statistical equivalence testing methods described by Fals-Stewart and Birchler (2002). Equivalence testing is used when the goal is to demonstrate that no significant differences exist between two conditions; in equivalence testing, the null hypothesis is that the two conditions differ by a significant amount. We used the confidence interval approach to examine whether the FT score recorded by UC-FT therapists was equivalent to the corresponding MDFT score. In this approach, an equivalence interval (EI) is defined as the mean of the reference group (MDFT fidelity sample) plus or minus 10%. Next, a confidence interval (CI) is defined by the following formula:

$$\mathrm{CI}_{90\%} = \mathrm{M}_{\mathrm{R}} - \mathrm{M}_{\mathrm{T}} +/- \mathrm{z}_{\alpha}\left(\mathrm{S}_{\mathrm{MR-MT}}\right)$$

In this equation M_R represents the average FT score for MDFT, M_T the average FT score for UC-FT, z_α the critical one-tailed value from the z distribution for the chosen value of α, and $S_{MR\text{-}MT}$ the pooled standard error. If the calculated CI falls within the EI, equivalence can be concluded. We then examined therapist differences in FT fidelity using two methods. First, we calculated the therapist-level ICC via an unconditional (no covariates) multi-level latent growth model using Mplus (Version 7; Muthén & Muthén, 2012). The ICC, a ratio of between-therapist variability to total variability in a given performance indicator, provides an estimate of therapist similarity that is independent of client effects in randomized studies (Adelson & Owen, 2012). The ICC is frequently used to estimate between-therapist differences in client outcomes within naturalistic treatment settings (e.g., Laska, Smith, Wislocki, Minami, & Wampold, 2013; Wiborg, Knoop, Wensing, & Bleijenberg, 2012). Second, SPC analyses (Deming, 1986) were conducted using SPSS (Version 23) to compare fidelity in UC-FT to the benchmark established by MDFT. Averaged FT scores for each UC-FT therapist were plotted on a control chart to check for outliers suggesting meaningful variation within the target sample compared to control limits (Callahan & Barisa, 2005). As described above (see *Benchmark Data Sources*), FT fidelity control limits were derived from an MDFT efficacy study (Hogue et al., 2006).

To examine whether UC-FT therapists achieved outcome benchmarks set in controlled trials of FT, we applied methods developed by Minami and colleagues (Minami, Serlin, Wampold, Kircher, & Brown, 2008) to benchmark treatment effectiveness for adult depression (Minami, Wampold, Serlin, Hamilton, & Brown, 2008) as well as MST for juvenile offenders (Curtis et al., 2009). The ES estimates recorded for Internalizing, Externalizing, and Delinquency outcomes were each benchmarked against the ES reported by Baldwin and colleagues (2012; see *Benchmark Data Sources*): $d = .21$. We first calculated ES estimate for each outcome:

$$d_{(i)} = \left(1 - 3/4n\ - 5\right)\ \mathrm{M}_{\mathrm{post}} - \mathrm{M}_{\mathrm{pre}}/\mathrm{SD}_{\mathrm{pre}}$$

We then tested whether the difference between each UC-FT ES and the benchmark ES exceeded Minami's statistical criterion for a clinically trivial difference (Δ = .02) using the noncentral *t* statistic (Minami, Serlin, et al., 2008). Calculations using algorithms provided by Minami (personal communication, 12.20.15) were completed using the statistical programming environment R (https://www.r-project.org/). To examine therapist differences in each outcome, we used the same two methods described above for fidelity variables: therapist-level ICC and SPC control charts. For the SPC analyses we charted simple change scores (6-month follow-up score minus Baseline score) averaged across clients for each therapist. Simple change scores are an efficient method of quantifying pre-post treatment change (Crits-Christoph et al., 2011). Because the outcome benchmark was a single point of data ($d = .21$) rather than an aggregate value with a mean and variance, it was not possible to construct upper and lower control limits using the benchmark ES. Instead, per SPC convention, upper and lower control limits for each outcome variable were derived from the plotted data and correspond to approximately three standard deviations above and below the given mean for the UC-FT sample (Noyez, 2009).

Results

Preliminary Analyses

Prior to conducting study analyses, variable transformations were applied to the delinquency outcome variable due to its skewed distribution (at baseline 13% of the sample reported zero delinquent acts in the past month). Exploratory analyses using graphical methods that depicted the effects of various (more and less) aggressive transformations suggested that a log transformation most appropriately addressed normality violations, reducing skewness (2.3 to 0.31) and kurtosis (6.0 to −0.21) at Baseline and also skewness (1.6 to 0.63) and kurtosis (2.5 to −0.78) at 6-month follow-up.

UC-FT Fidelity: Benchmarking

Equivalence testing was first used to compare the raw (unadjusted) FT adherence score for UC-FT (M = 2.7; SD = 0.53) to the score earned by MDFT therapists in an efficacy trial (M = 2.0; SD = 0.35). Using the 10% criteria for interval width, the equivalence interval defined by the benchmark MDFT data was: 1.8 to 2.2. The 90% confidence interval was then calculated: −0.57 to −0.83. Because this 90% CI range does not overlap with the defined range of the equivalence interval, it can be concluded that the raw FT score for UC-FT is not statistically equivalent to the FT score for MDFT—that is, the UC-FT score is statistically larger than the MDFT score. We then tested the inflation-adjusted FT adherence score (M = 2.3; SD = 0.70) to the same equivalence interval for the MDFT benchmark score. The 90% confidence interval was re-calculated for the inflation-adjusted mean: −0.30 to −0.34. As with the unadjusted score, the 90% CI for the inflation-adjusted FT adherence score does not overlap with the defined range of the equivalence interval. Thus, even after adjusting the UC-FT score for inflation, it remained significantly higher than the MDFT benchmark.

UC-FT Fidelity: Therapist Differences

Multilevel modeling examining therapist heterogeneity in FT adherence scores indicated that 50% (ICC = .50) of variance in the FT score was due to between-therapist differences.

Figure 1 depicts the SPC chart containing the FT score for each of the 13 UC-FT therapists. Each point on the chart represents the average FT score across all clients seen by a particular therapist. The chart also depicts the inflation-adjusted mean FT score for the UC-FT sample (indicated by the solid line) as well as upper and lower control limits (labeled MDFT UCL and MDFT LCL) that are based on criterion values (i.e., benchmarks) derived from the MDFT efficacy trial. As reported previously, the observed inflation-adjusted mean UC-FT score was 2.3 and the benchmark mean from the MDFT trial was 2.0. As shown in Figure 1, all UC-FT scores fall within the upper and lower control limits set in the MDFT trial. These results indicate that each of the 13 therapists in this community sample met the FT fidelity bandwidth standards established in a controlled trial.

UC-FT Outcomes: Benchmarking

UC-FT client changes in Internalizing, Externalizing, and Delinquency from baseline to 6-month follow-up were benchmarked against aggregate client changes reported in controlled FT trials as operationalized by Baldwin and colleagues (2012) using a standardized ES: $d = 0.21$. Standardized ES estimates were calculated for UC-FT clients using Minami's method (Minami, Serlin, et al., 2008); results were $d = 0.31$ for Internalizing, $d = 0.51$ for Externalizing, and $d = 0.87$ for Delinquency. The UC-FT ES estimates were compared to the benchmark ES using the non-central *t* test. For Internalizing, UC-FT outcomes were found to be clinically equivalent to the benchmark ($t(28) = 1.67$, $\lambda = 1.02$, $p = 0.27$). UC-FT outcomes significantly exceeded the benchmark for both Externalizing ($t(29) = 2.79$, $\lambda = 1.04$, $p = 0.05$) and Delinquency ($t(28) = 4.69$, $\lambda = 1.02$, $p < 0.01$), indicating that client change in the study sample was larger in magnitude for these outcomes than client change in the benchmark sample.

UC-FT Outcomes: Therapist Differences

Multilevel latent growth curve modeling revealed that 2% (ICC = .02) of variability in baseline Delinquency scores and 7% (ICC = .07) of variability at 6-month follow-up were due to between-therapist differences. For Externalizing, ICCs were .03 for baseline and .02 for followup; for Internalizing, ICCs were .02 for baseline and .01 for follow-up.

SPC charts containing change scores on each outcome are depicted in Figures 2–4. As described previously, upper and lower control limits were derived from the sample data for each outcome. Thus, SPC outcome charts depict variability among individual therapists in client outcomes, but they do not represent to the extent to which UC-FT therapists individually varied in achieving the benchmark ES for client change ($d = 0.21$). In each SPC outcome chart, client change is plotted for each of the 12 UC-FT therapists for whom complete outcome data were available, identified by therapist ID numbers along the x-axis to permit cross-chart comparisons. Points on the chart represent change scores averaged across all clients seen by a particular therapist. The solid line represents the sample mean change score, and the dashed lines represent the upper and lower control limits (three standard deviations above and below the mean).

Figure 2 depicts the SPC chart for Internalizing. The average change was 2.5 (SD = 5.2). While all points fall within the upper and lower control limits, there appears to be considerable variability across therapists. In particular, therapists 30 and 32 appear close to the lower control limit and thus out of sync with other therapists. Figure 3 depicts the chart for Externalizing. The average change was 5.3 (SD = 6.1), and all points fall within the upper and lower control limits. However, therapist 30 is approaching the lower control limit. Figure 4 depicts the chart for Delinquency, with a mean change of 0.31 (SD = 0.32). Here, all points fall within the upper and lower control limits, and all appear relatively close to the mean.

Discussion

Study results showed that community family therapists treating adolescent behavior problems in usual care successfully achieved performance benchmarks established in controlled studies of manualized FT. UC-FT therapists were statistically superior to the fidelity benchmark for adherence to core FT techniques, even after adjusting for therapist-report score inflation. Regarding change in client functioning at 6-month follow-up, UC-FT therapists were statistically equivalent to the benchmark for internalizing symptoms and superior for externalizing symptoms and delinquent acts. Contrary to hypotheses about robust individual therapist differences, UC-FT clinicians demonstrated a high degree of performance uniformity: Each one approximated the fidelity benchmark, and only 2 of 12 produced outcomes that appeared relatively weak on any of the client change indicators.

The fidelity results confirm findings from a previous exploratory study on adherence to core FT techniques conducted at the same treatment site on an earlier cohort of therapists (Hogue & Dauber, 2013). Note that even with inflation correction, the adjusted UC-FT mean adherence score (2.3) still significantly exceeded the observer-reported MDFT benchmark (2.0). Of course it is difficult for other evaluators to calculate valid inflation-adjustment formulae for their given measures and samples, absent the advantage of having collateral fidelity ratings by therapists and observers on the same set of treatment sessions. As research on treatment implementation in UC advances, it should be possible for widely used therapist-report fidelity measures to develop tool-specific inflation-adjustment indices, allowing providers and evaluators to apply the correction to any sample assessed with those measures.

Regarding client outcomes, to our knowledge this is the first study to benchmark treatment effects for community therapists delivering non-manualized FT for adolescent behavior problems, and the positive results align with findings by Curtis and colleagues (2009) for community therapists trained in MST. Of course these results do not support the contention that core element FT implemented in everyday practice is fundamentally equivalent to manualized FT implemented with extramural training and monitoring by EBT purveyor organizations (see Henggeler & Sheidow, 2012); such comparisons require randomized controlled investigation. Even so it remains noteworthy that mainstream FT practitioners working in routine conditions (see Hoagwood, 2005) can yield measurable successes in FT fidelity and outcome.

The therapist-report raw mean score for FT adherence (2.7) and the inflation-adjusted mean (2.3) both fall between the scale anchor values of 2 (*A little bit*) and 3 (*Moderately*); the mean score for the benchmark MDFT sample was 2.0, equal to the anchor of 2 (*A little bit*). These adherence levels are consistent with levels reported in previous observational fidelity studies across a range of manualized treatment approaches and populations (e.g., Carroll et al., 2000; Hill, O’Grady, & Elkin, 1992; Hogue et al., 2008) and can be considered the roughly “normal” level of averaged ratings for multi-item fidelity scales. These below-midpoint mean scores likely reflect the fact that neither research-hired nor UC clinicians can be expected to deliver a full roster of discrete techniques from the governing treatment model in any one session, in light of prevailing time and client tolerance limits. Indeed, an active therapist can implement one or two interventions very thoroughly during a given session yet still receive a below-midpoint mean adherence score that has been averaged across multiple scale items. Another metric for judging the density of EBT delivery in usual care might be tabulating the proportion of sessions in which one (or a few) discrete techniques are scored at or above the midpoint value, indicating the presence of considerable/extensive EBT activity (e.g., Southam-Gerow et al., 2016); though beyond the scope of the current study, analyses of this kind would further enrich our understanding of UC treatment processes (Hurlburt et al., 2010).

As a secondary aim, this study demonstrated the value added by SPC analyses for diagnosing therapist heterogeneity, over and above conventional estimation via ICCs. The between-therapistICC values at pre- and post-treatment for the three client outcomes ranged from .01 to .07, which falls squarely in the range typically reported in the behavioral treatment literature (Adelson & Owen, 2012). The heterogeneitycoefficient for FT fidelity was substantially higher: ICC = .50. Given that individual therapist differences in intervention fidelity have not yet been mapped, it is impossible to surmise whether this coefficient represents the norm for implementation effects—because therapists have proximal control over treatment processes but only distal control over client outcomes?—or is simply a sample aberration. In any event, raw ICC values for fidelity and outcome offer no insight into which individual therapists might be thriving or failing, nor whether the overall spread is homogenous or instead punctuated by relative outliers on the high and/or low end of the performance spectrum.

As evidenced here, SPC control charts offer this kind of articulated information on therapist performance differences in easily digested graphics. In this particular sample, SPC analysis disconfirmed any impression made by the high ICC for fidelity (.50) that UC-FT therapists were widely discrepant from one another in FT adherence; on the contrary, no individual therapist appeared as a fidelity outlier. Certainly the absence of notable deviations among study therapists for either fidelity or outcome indicators may be function of the small number of clinicians and clients sampled (see *Study Strengths and Limitations*). To wit: Is it justified to single out therapist 30 due to his/her lower performance on two outcomes averaged across two clients? SPC is most powerful when employed in an assembly-line context as part of a continuous production process (e.g., Dey, Sluyter, & Keating, 1994), for example tracking clinical staff within a provider system over several years, in which deviations can be detected for a given time window as well as a given therapist (Green, 1999). And because it is designed to distinguish between normal variation and systematic uncontrolled variation, SPC may be ideal for supporting line clinicians who are responsible for consistent delivery of EBTs yet also expected to show natural variation in implementation across sessions and caseloads (Delgadillo et al., 2014). Inexpensive and user-friendly SPC methods can be readily merged into clinical data management systems that guide decision making about how treatment is progressing and when corrective action might be needed (e.g., Chorpita et al., 2008), even in agencies with rudimentary computing resources (for example, SPC analyses are supported in Excel).

Study Strengths and Limitations

The main strength of this benchmarking study was its comprehensive assessment design. Treatment fidelity was evaluated along with client outcomes, creating the opportunity to benchmark both inputs and outputs of behavioral treatment. Adherence to specific treatment techniques was assessed at virtually every session using an observationally validated therapist-report fidelity measure (Hogue, Dauber, Lichvar, et al., 2015), creating a dense set of fidelity data for analysis. Outcomes included dimensional measures of both internalizing and externalizing problems that are prevalent among adolescent referrals to behavioral care. Another strength was sample diversity: Participants were primarily Hispanic and African American, were balanced between male and female, and presented with an array of conduct, mood and anxiety, and substance use disorders. The demographic and clinical profile of study participants, an “unmet need” sample recruited from a network of school and community referral sources in the same catchment area as the treatment site, was a comfortable match with the profile of the site’s existing referral stream. Yet there may be important differences between the study sample and the usual referral stream (e.g., symptom severity, treatment history) that are ultimately related to treatment fidelity and outcome; because the site did not collect standardized intake or outcome data on non-study cases, potential differences are unknown.

The main study limitation was the sampling design. There was a relatively small number of therapists and cases, all from a single treatment site. Thus therapist participants are not broadly representative of the family therapy workforce, making it impossible to conclude that their strong overall performance is generalizable to other front-line family therapists. As importantly, the small sample size limits the generalizability of conclusions to be drawn about observed therapist differences on performance indicators. The current study could not control for client-level variables due to insufficient power (generally a ratio of 5–10 clients per therapist is recommended; see Adelson & Owen, 2012; Erickson, Tonigan, & Winhusen, 2012. Absent the capacity to model client effects, it is impossible to determine whether observed therapist differences are driven primarily by differences among therapists, differences among clients, or a combination of the two. Indeed, a previous well-powered study of client effects on outcomes in the parent trial (Hogue, Henderson, & Schmidt, 2016) found that multiple baseline client characteristics—including demographic (e.g., age), clinical (e.g., symptom severity), and developmental psychopathology indicators (e.g., depression, delinquent peer affiliations)—significantly predicted change in delinquency and substance use at one-year follow-up.

The small sample size also limits the generalizability of the benchmarking analyses, in that it provided a narrow opportunity to observe therapists with markedly subpar performances that would stand out during more expansive SPC analyses. Of 12 therapists with client outcome data, 3 treated one case only; if that case happened to be a performance outlier within that therapist’s population of cases, the results logged in this study would be non-representative. More generally, larger proportions of sample therapists with one case only yield less precise estimates of therapist effect correlations, which are governed by within-therapist variance. Also, a bigger sample would have allowed us to test whether key therapist characteristics (e.g., demographics, experience, therapeutic orientation) predicted strong versus weak performance (Bearman et al., 2013; Beidas & Kendall, 2010; Blatt, Sanislow, Zuroff, & Pilkonis, 1996). Future studies might also benefit from additional specificity in outcome benchmarking, that is, having a separate gold standard for each unique outcome.

Study Implications: Developing Therapist-Report Fidelity Measures for Quality Assurance

This study featured reliable therapist reports of fidelity to family therapy techniques. This was an unusual luxury. Therapists have demonstrated uniformly poor concordance with non-participant observers when rating fidelity to treatment techniques, even in studies that directly retrofitted validated observational measures of EBT fidelity for use as self-report tools (reviwed in Hogue, Dauber, Lichvar, et al., 2015). Yet there remains hope for salvaging the reliability of therapist-report fidelity: train front-line clinicians to be fluent in self-rating, using procedures analogous to those used with observational coders (Hurlburt et al., 2010). These procedures include adapting observational measures to capture the desired treatment techniques on self-report scales, training clinicians to self-rate via didactic instruction and in vivo practice guided by experts, and periodic monitoring of self-report data via peer-supported review of ratings coupled with retraining for items/clinicians with declining reliability (see Hill, 1991).

Is such a rigorous approach feasible in everyday practice settings? There are reasons to believe so. EBT purveyors with existing observational fidelity measures are primed to fashion appropriate self-report tools and training procedures, and the initial training of line clinicians in self-report reliability appears to fit snugly within the broader goals of EBT dissemination, QA, and sustainment (Hogue et al., 2013). It may not be prohibitively difficult to train community therapists to be reliable fidelity self-raters, given that several observational studies employed practicing therapists as coders (e.g., Hogue et al., 2008). One caveat is that therapist-report measures cannot legitimately capture the quality (i.e., competence) of treatment implementation (Barber et al., 2007), though treatment adherence data themselves can and should play a pivotal role in broader judgments about service quality (McLeod et al., 2013). Until proven otherwise, assessment of treatment quality remains the province of supervisors and other observers. Another caveat (discussed above) is the tendency for score inflation among therapist reporters.

Study Implications: Correcting (Unwanted) Therapist Differences during Quality Improvement

The emerging healthcare market, spurred by the Affordable Care Act, is focused on increasing quality and accountability in behavioral care using the complementary procedural linchpins of QA procedures plus quality improvement (QI) procedures, in which performance data are used to formulate and implement plans to systematically improve EBT implementation (see Hoagwood, 2013). Systematic monitoring of various types of therapist performance data— including treatment fidelity (McLeod et al., 2013) and client outcomes —is a key aspect of the QA/QI process. The SPC procedures described in this study can provide unique insights on therapist differences in clinical performance, especially if there is a large enough performance sample to reliably diagnose when the “effect” belongs primarily to the therapist (which was not possible in the current study). For example, if SPC analyses showed that particular therapist(s) placed well outside research-based control limits for EBT fidelity, this would suggest that the best means to improve outcomes is to enhance fidelity levels by strengthening clinician training and QA protocols. In contrast, if therapists were basically adherent to fidelity control limits but still missed the mark in outcome success, this may direct QI efforts toward adapting/enhancing the EBT itself to improve its potency in community settings—or replacing it altogether. In these ways, monitoring therapist effects on a continuous basis over a sizable performance sample can produce a call to action for correcting clinician performance by means of appropriate shaping of local QA and QI practices.

Acknowledgments

Preparation of this article was supported by grant R01DA023945 from the National Institute on Drug Abuse. The authors are extremely grateful to Takuya Minami for providing technical assistance in completing benchmarking analyses. We acknowledge the dedicated contributions of Molly Bobek, Jaqueline Horan Fisher, Candace Johnson, and Emily Lichvar in supporting this work. We also thank the therapists working at the Roberto Clemente Center who generously agreed to submit regular self-reports of therapeutic interventions in the hopes of advancing the clinical science in their field.

References

Achenbach, TM., Rescorla, LA. ASEBA School Age Forms and Profiles. Burlington, Vt.: ASEBA; 2001.

Adelson JL, Owen J. Bringing the psychotherapist back: Basic concepts for reading articles examining therapist effects using multilevel modeling. Psychotherapy. 2012; 49:152. [PubMed: 21967075]

American Psychiatric Associatio. Diagnostic and statistical manual of mental disorders. 4th. Washington, DC: APA; 2000. text revision

Baldwin SA, Christian S, Berkeljon A, Shadish W, Bean R. The effects of family therapies for adolescent delinquency and substance abuse: A meta-analysis. Journal of Marital and Family Therapy. 2012; 38:281–304. [PubMed: 22283391]

Barber JP, Sharpless B, Klostermann S, McCarthy KS. Assessing intervention competence and its relation to therapy outcome: A selected review derived from the outcome literature. Professional Psychology: Research and Practice. 2007; 38:493–500.

Barth RP, Kolivoski KM, Lindsey MA, Lee BR, Collins KS. Translating the common elements approach: Social work's experiences in education, practice, and research. Journal of Clinical Child & Adolescent Psychology. 2014; 43(2):301–311. [PubMed: 24245958]

Bearman SK, Weisz JR, Chorpita BF, Hoagwood K, Ward A, Research Network on Youth Mental Health. More practice, less preach? The role of supervision processes and therapist characteristics in EBP implementation. Administration and Policy in Mental Health and Mental Health Services Research. 2013; 40:518–529. [PubMed: 23525895]

Beidas RS, Kendall PC. Training therapists in evidence‐based practice: A critical review of studies from a systems‐contextual perspective. Clinical Psychology: Science and Practice. 2010; 17:1–30. [PubMed: 20877441]

Blatt SJ, Sanislow CA III, Zuroff DC, Pilkonis PA. Characteristics of effective therapists: Further analyses of data from the National Institute of Mental Health Treatment of Depression Collaborative Research Program. Journal of Consulting and Clinical psychology. 1996; 64:1276. [PubMed: 8991314]

Bond GR, Becker DR, Drake RE. Measurement of fidelity of implementation of evidence-based practices: Case example of the IPS Fidelity Scale. Clinical Psychology: Science and Practice. 2011; 18:126–141.

Brosan L, Reynolds S, Moore RG. Self-evaluation of cognitive therapy performance: Do therapists know how competent they are? Behavioural and Cognitive Psychotherapy. 2008; 36:581–587.

Callahan CD, Barisa MT. Statistical process control and rehabilitation outcome: The single-subject design reconsidered. Rehabilitation Psychology. 2005; 50:24–33.

Carroll KM, Nich C, Rounsaville BJ. Utility of therapist session checklists to monitor delivery of coping skills treatment for cocaine abusers. Psychotherapy Research. 1998; 8:307–320.

Carroll KM, Nich C, Sifry R, Nuro KF, Frankforter TL, et al. A general system for evaluating therapist adherence and competence in psychotherapy research in the addictions. Drug & Alcohol Dependence. 2000; 57:225–238. [PubMed: 10661673]

Chorpita BF, Becker KD, Daleiden EL. Understanding the common elements of evidence-based practice: Misconceptions and clinical examples. Journal of the American Academy of Child & Adolescent Psychiatry. 2007; 46:647–652. [PubMed: 17450056]

Chorpita BF, Bernstein A, Daleiden EL. Driving with roadmaps and dashboards: Using information resources to structure the decision models in service organizations. Administration and Policy in Mental Health and Mental Health Services Research. 2008; 35:114–123. [PubMed: 17987376]

Chorpita B, Daleiden E, Ebesutani C, Young J, Becker K, Nakamura B, et al. Evidence-based treatments for children and adolescents: An updated review of indicators of efficacy and effectiveness. Clinical Psychology: Science and Practice. 2011; 18:154–172.

Cohen, J. Statistical power analysis for the behavioral sciences. 2nd. Hillsdale, NJ: Erlbaum; 1988.

Crits-Christoph P, Gibbons MBC, Hamilton J, Ring-Kurtz S, Gallop R. The dependability of alliance assessments: The alliance-outcome correlation is larger than you might think. Journal of Consulting and Clinical Psychology. 2011; 79:267–278. [PubMed: 21639607]

Crits-Christoph P, Mintz J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. Journal of Consulting and Clinical Psychology. 1991; 59:20. [PubMed: 2002139]

Curran PJ, Hussong AM. Integrative data analysis: The simultaneous analysis of multiple data sets. Psychological Methods. 2009; 14:81–100. [PubMed: 19485623]

Curtis NM, Ronan KR, Heiblum N, Crellin K. Dissemination and effectiveness of multisystemic treatment in New Zealand: a benchmarking study. Journal of Family Psychology. 2009; 23:119. [PubMed: 19364207]

Delgadillo J, McMillan D, Leach C, Lucock M, Gilbody S, Wood N. Benchmarking routine psychological services: a discussion of challenges and methods. Behavioural and Cognitive Psychotherapy. 2014; 42:16–30. [PubMed: 23092729]

Deming, WE. Out of the Crisis. Cambridge, MA: MIT Press; 1986.

Dey ML, Sluyter GV, Keating JE. Statistical process control and direct care staff performance. Journal of Mental Health Administration. 1994; 21:201–209. [PubMed: 10145995]

Elkin I. A major dilemma in psychotherapy outcome research: Disentangling therapists from therapies. Clinical Psychology: Science and Practice. 1999; 6:10–32.

Elliott, DS., Huizinga, D., Ageton, SS. Explaining delinquency and drug use. Beverly Hills, CA: Sage Publications; 1985.

Erickson SJ, Tonigan JS, Winhusen T. Therapist effects in a NIDA CTN intervention trial with pregnant substance abusing women: Findings from a RCT with MET and TAU conditions. Alcoholism Treatment Quarterly. 2012; 30:224–237.

Fals-Stewart W, Birchler GR. Behavioral couples therapy with alcoholic men and their intimate partners: The comparative effectiveness of bachelor’s and master’s level counselors. Behavior Therapy. 2002; 33:123–147.

Flicker SM, Waldron HB, Turner CW, Brody JL, Hops H. Ethnic matching and treatment outcome with Hispanic and Anglo substance-abusing adolescents in family therapy. Journal of Family Psychology. 2008; 22:439. [PubMed: 18540772]

Garland AF, Bickman L, Chorpita BF. Change what? Identifying quality improvement targets by investigating usual mental health care. Administration and Policy in Mental Health and Mental Health Services Research. 2010; 37:15–26. [PubMed: 20177769]

Garland AF, Haine-Schlagel R, Brookman-Frazee L, Baker-Ericzen M, Trask E, Fawley-King K. Improving community-based mental health care for children: Translating knowledge into action. Administration and Policy in Mental Health and Mental Health Services Research. 2013; 40:6–22. [PubMed: 23212902]

Garland AF, Hawley KM, Brookman-Frazee LI, Hurlburt M. Identifying common, core elements of evidence-based practice for children with disruptive behavior disorder. Journal of the American Academy of Child & Adolescent Psychiatry. 2008; 47:505–514. [PubMed: 18356768]

Green RS. The application of statistical process control to manage global client outcomes in behavioral healthcare. Evaluation and Program Planning. 1999; 22:199–210. [PubMed: 24011413]

Haley, J. Problem-solving therapy. San Francisco: Jossey-Bass; 1987.

Hedges, LV., Olkin, I. Statistical methods for meta-analysis. Orlando, FL: Academic; 1985.

Henggeler SW, Pickrel SG, Brondino MJ. Multisystemic treatment of substance-abusing and-dependent delinquents: Outcomes, treatment fidelity, and transportability. Mental Health Services Research. 1999; 1:171–184. [PubMed: 11258740]

Henggeler SW, Sheidow AJ. Empirically supported family-based treatments for conduct disorder and delinquency in adolescents. Journal of Marital and Family Therapy. 2012; 38:30–58. [PubMed: 22283380]

Hill, CE. Almost everything you ever wanted to know about how to do process research on counseling and psychotherapy but didn’t know who to ask. In: Hill, CE., Schneider, LJ., editors. Research in Counseling. Hillsdale, NJ: Erlbaum; 1991. p. 85-118.

Hill CE, O’Grady KE, Elkin I. Applying the Collaborative Study Psychotherapy Rating Scale to rate therapist adherence in cognitive-behavior therapy, interpersonal therapy, and clinical management. Journal of Consulting and Clinical Psychology. 1992; 60:73–79. [PubMed: 1556289]

Hoagwood KE. Family-based services in children’s mental health: A research review and synthesis. Journal of Child Psychology and Psychiatry. 2005; 46:690–713. [PubMed: 15972066]

Hoagwood KE. Don’t mourn: Organize. Reviving mental health services research for healthcare quality improvement. Clinical Psychology: Science and Practice. 2013; 20:120–126.

Hogue A, Dauber S. Assessing fidelity to evidence-based practices in usual care: The example of family therapy for adolescent behavior problems. Evaluation and Program Planning. 2013; 37:21–30. [PubMed: 23314000]

Hogue A, Dauber S, Chinchilla P, Fried A, Henderson CE, et al. Assessing fidelity in individual and family therapy for adolescent substance abuse. Journal of Substance Abuse Treatment. 2008; 35:137–147. [PubMed: 17997268]

Hogue A, Dauber S, Henderson CE. Therapist self-report of evidence-based practices in usual care for adolescent behavior problems: Factor and construct validity. Administration and Policy in Mental Health and Mental Health Services Research. 2014; 41:126–139. [PubMed: 23124275]

Hogue A, Dauber S, Henderson CE, Bobek M, Johnson C, Lichvar E, Morgenstern J. Randomized trial of family therapy versus non-family treatment for adolescent behavior problems in usual care. Journal of Clinical Child and Adolescent Psychology. 2015; 44:954–969. [PubMed: 25496283]

Hogue A, Dauber S, Henderson CE, Liddle HA. Reliability of therapist self-report on treatment targets and focus in family-based intervention. Administration and Policy in Mental Health and Mental Health Services Research. 2014; 41:697–705. [PubMed: 24068479]

Hogue A, Dauber S, Lichvar E, Bobek M, Henderson CE. Validity of therapist self-report ratings of fidelity to evidence-based practices for adolescent behavior problems: Correspondence between therapists and observers. Administration and Policy in Mental Health and Mental Health Services Research. 2015; 42:229–243. [PubMed: 24711046]

Hogue A, Dauber S, Samuolis J, Liddle HA. Treatment techniques and outcomes in multidimensional family therapy for adolescent behavior problems. Journal of Family Psychology. 2006; 20:535–543. [PubMed: 17176187]

Hogue A, Henderson CE, Ozechowski TJ, Robbins MS. Evidence base on outpatient behavioral treatments for adolescent substance use: Updates and recommendations 2007–2013. Journal of Clinical Child and Adolescent Psychology. 2014; 43:697–720.

Hogue A, Henderson CE, Schmidt AT. Multidimensional predictors of treatment outcome in usual care for adolescent conduct problems and substance use. Administration and Policy in Mental Health and Mental Health Services Research. 2016; doi: 10.1007/s10488-016-0724-7

Hogue A, Liddle HA. Family-based treatment for adolescent substance abuse: Controlled trials and new horizons in services research. Journal of Family Therapy. 2009; 31:126–154. [PubMed: 21113237]

Hogue A, Ozechowski TJ, Robbins MR, Waldron HB. Making fidelity an intramural game: Localizing quality assurance procedures to promote sustainability of evidence-based practices in usual care. Clinical Psychology: Science and Practice. 2013; 20:60–77.

Hoyer RW, Ellis WC. A graphical exploration of SPC: Part 1-SPC's definitions and procedures. Quality Progress. 1996; 29:65–72.

Hunsley J, Lee CM. Research-informed benchmarks for psychological treatments: Efficacy studies, effectiveness studies, and beyond. Professional Psychology: Research and Practice. 2007; 38:21.

Hurlburt MS, Garland AF, Nguyen K, Brookman-Frazee L. Child and family therapy process: Concordance of therapist and observational perspectives. Administration and Policy in Mental Health and Mental Health Services Research. 2010; 37:230–244. [PubMed: 19902347]

Institute of Medicine. Improving the quality of healthcare for mental and substance-use conditions. Washington, D.C.: National Academy Press; 2006.

Laska KM, Smith TL, Wislocki AP, Minami T, Wampold BE. Uniformity of evidence-based treatments in practice? Therapist effects in the delivery of cognitive processing therapy for PTSD. Journal of Counseling Psychology. 2013; 60:31. [PubMed: 23356465]

Löfholm CA, Eichas K, Sundell K. The Swedish implementation of multisystemic therapy for adolescents: Does treatment experience predict treatment adherence? Journal of Clinical Child & Adolescent Psychology. 2014; 43:643–655. [PubMed: 24661234]

Martino S, Ball S, Nich C, Frankforter TL, Carroll KM. Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. Psychotherapy Research. 2009; 19:181–193. [PubMed: 19396649]

McHugh R, Barlow DH. The dissemination and implementation of evidence-based psychological treatment: A review of current efforts. American Psychologist. 2010; 65:73–84. [PubMed: 20141263]

McLeod BD, Southam-Gerow MA, Tully CB, Rodríguez A, Smith MM. Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. Clinical Psychology: Science and Practice. 2013; 20:14–32. [PubMed: 23935254]

Minami T, Serlin RC, Wampold BE, Kircher JC, Brown GS. Using clinical trials to benchmark effects produced in clinical practice. Quality and Quantity. 2008; 42:513–525.

Minami T, Wampold BE, Serlin RC, Hamilton EG, Brown GS, Kircher JC. Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. Journal of Consulting and Clinical Psychology. 2008; 76:116–124. [PubMed: 18229989]

Minuchin, S., Fishman, HC. Family therapy techniques. Cambridge, MA: Harvard University Press; 1981.

Muthen, B., Muthen, L. Mplus User's Guide. Los Angeles, CA: Muthen & Muthen; 2012.

Noyez L. Control charts, cusum techniques, and funnel plots: A review of methods for monitoring performance in healthcare. Interactive Cardiovascular and Thoracic Surgery. 2009; 9:494–499. [PubMed: 19509097]

Ozechowski TJ, Waldron HB. Assertive outreach strategies for narrowing the adolescent substance abuse treatment gap: Implications for research, practice, and policy. Journal of Behavioral Health Services & Research. 2010; 37:40–63. [PubMed: 18690540]

Project MATCH Research Group. Therapist effects in three treatments for alcohol problems. Psychotherapy Research. 1998; 8:455–474.

Schoenwald SK. It's a bird, It's a plane, It's… fidelity measurement in the real world. Clinical Psychology: Science and Practice. 2011; 18:142–147. [PubMed: 21691439]

Schoenwald SK, Carter RE, Chapman JE, Sheidow AJ. Therapist adherence and organizational effects on change in youth behavior problems one year after multisystemic therapy. Administration and Policy in Mental Health and Mental Health Services Research. 2008; 35:379–394. [PubMed: 18561019]

Schoenwald SK, Chapman JE, Sheidow AJ, Carter RE. Long-term youth criminal outcomes in MST transport: The impact of therapist adherence and organizational climate and structure. Journal of Clinical Child & Adolescent Psychology. 2009; 38:91–105. [PubMed: 19130360]

Schoenwald SK, Henggeler SW, Brondino MJ, Rowland MD. Multisystemic therapy: Monitoring treatment fidelity. Family Process. 2000; 39:83–103. [PubMed: 10742933]

Schoenwald SK, Letourneau EJ, Halliday-Boykins C. Predicting therapist adherence to a transported family-based treatment for youth. Journal of Clinical Child and Adolescent Psychology. 2005; 34(4):658–670. [PubMed: 16232063]

Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini International Neuropsychiatric Interview (MINI): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. Journal of Clinical Psychiatry. 1998; 59:22–33.

Southam-Gerow MA, McLeod BD, Arnold CC, Rodríguez A, Cox JR, Reise SP, Kendall PC. Initial development of a treatment adherence measure for cognitive–behavioral therapy for child anxiety. Psychological Assessment. 2016; 28:70–80. [PubMed: 26011477]

Spilka MJ, Dobson KS. Promoting the Internationalization of Evidence‐Based Practice: Benchmarking as a Strategy to Evaluate Culturally Transported Psychological Treatments. Clinical Psychology: Science and Practice. 2015; 22:58–75.

Tanner-Smith EE, Wilson SJ, Lipsey MW. The comparative effectiveness of outpatient treatment for adolescent substance abuse: A meta-analysis. Journal of Substance Abuse Treatment. 2012; 44:145–158. [PubMed: 22763198]

Wampold BE, Serlin RC. The consequence of ignoring a nested factor on measures of effect size in analysis of variance. Psychological Methods. 2000; 5:425. [PubMed: 11194206]

Weersing VR. Benchmarking the effectiveness of psychotherapy: Program evaluation as a component of evidence-based practice. Journal of the American Academy of Child & Adolescent Psychiatry. 2005; 44:1058–1062. [PubMed: 16175111]

Weersing RV, Weisz JR. Community clinic treatment of depressed youth: Benchmarking usual care against CBT clinical trials. Journal of Consulting and Clinical Psychology. 2002; 70:299–310. [PubMed: 11952188]

Weersing RV, Weisz JR, Donenberg GR. Development of the Therapy Procedures Checklist: A therapist-report measure of technique use in child and adolescent treatment. Journal of Clinical Child Psychology. 2002; 31:168–80.

Weisz JR, Chorpita BF, Palinkas LA, Schoenwald SK, Miranda J, Bearman SK, et al. Testing standard and modular designs for psychotherapy treating depression, anxiety, and conduct problems in youth: A randomized effectiveness trial. Archives of General Psychiatry. 2012; 69:274–282. [PubMed: 22065252]

Wiborg JF, Knoop H, Wensing M, Bleijenberg G. Therapist effects and the dissemination of cognitive behavior therapy for chronic fatigue syndrome in community-based mental health care. Behaviour Research and Therapy. 2012; 50:393–396. [PubMed: 22504122]

Zucker DM. An analysis of variance pitfall: The fixed effects analysis in a nested design. Educational and Psychological Measurement. 1990; 50:731–738.

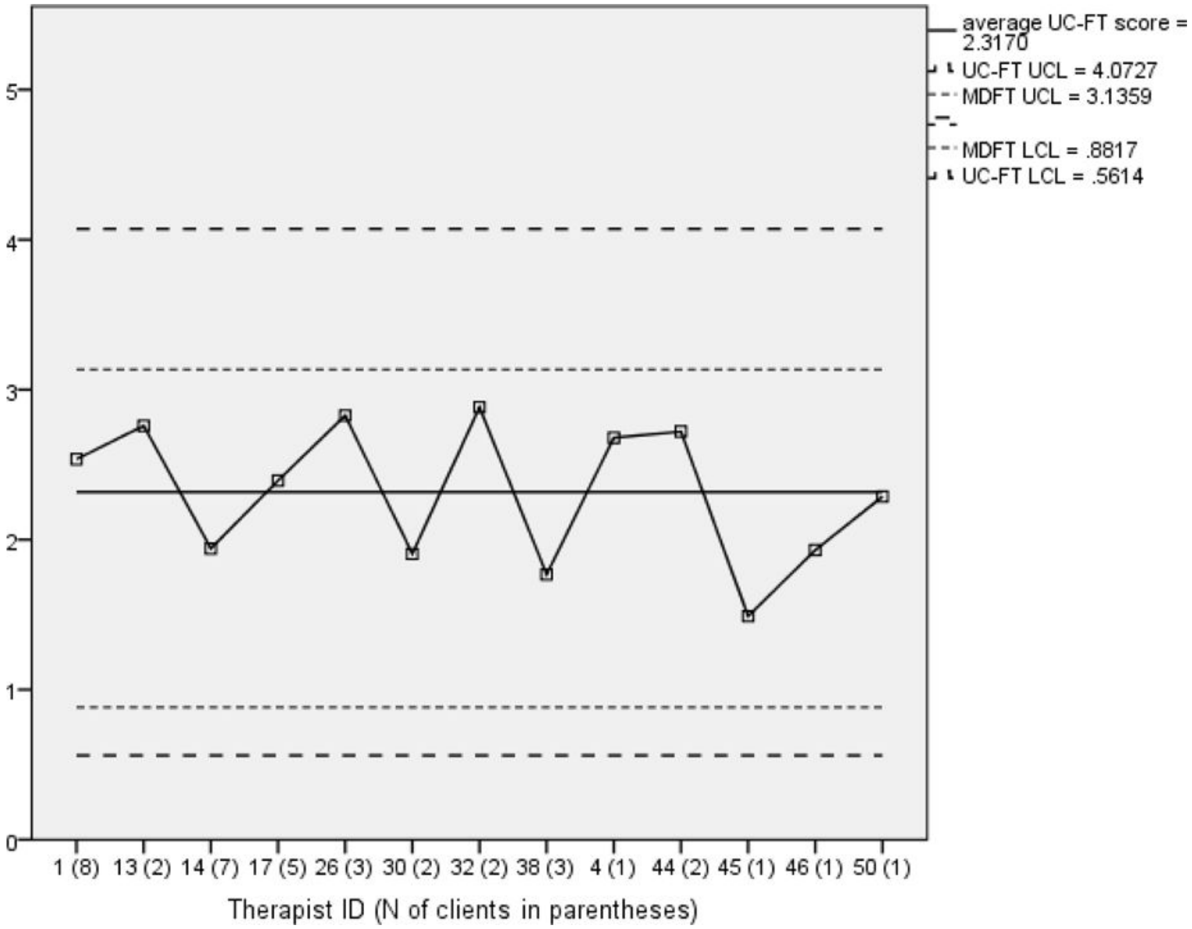

Figure 1.

SPC mean chart for UC-FT therapists benchmarked against MDFT efficacy data: Inflation-adjusted therapist-report fidelity to core family therapy treatment techniques.

Note. The numbers on the x-axis correspond to the Therapist identification code for each of the 13 therapists, with the number of clients treated by each therapist in parentheses. Points on the chart are averages of FT scores across clients for each therapist, adjusted for inflation. UC-FT UCL = Usual Care Family Therapy Upper Control Limit; MDFT UCL = Multidimensional Family Therapy Upper Control Limit (specified for benchmarking purposes); UC-FT LCL = Usual Care Family Therapy Lower Control Limit; MDFT LCL = Multidimensional Family Therapy Lower Control Limit (specified for benchmarking purposes).

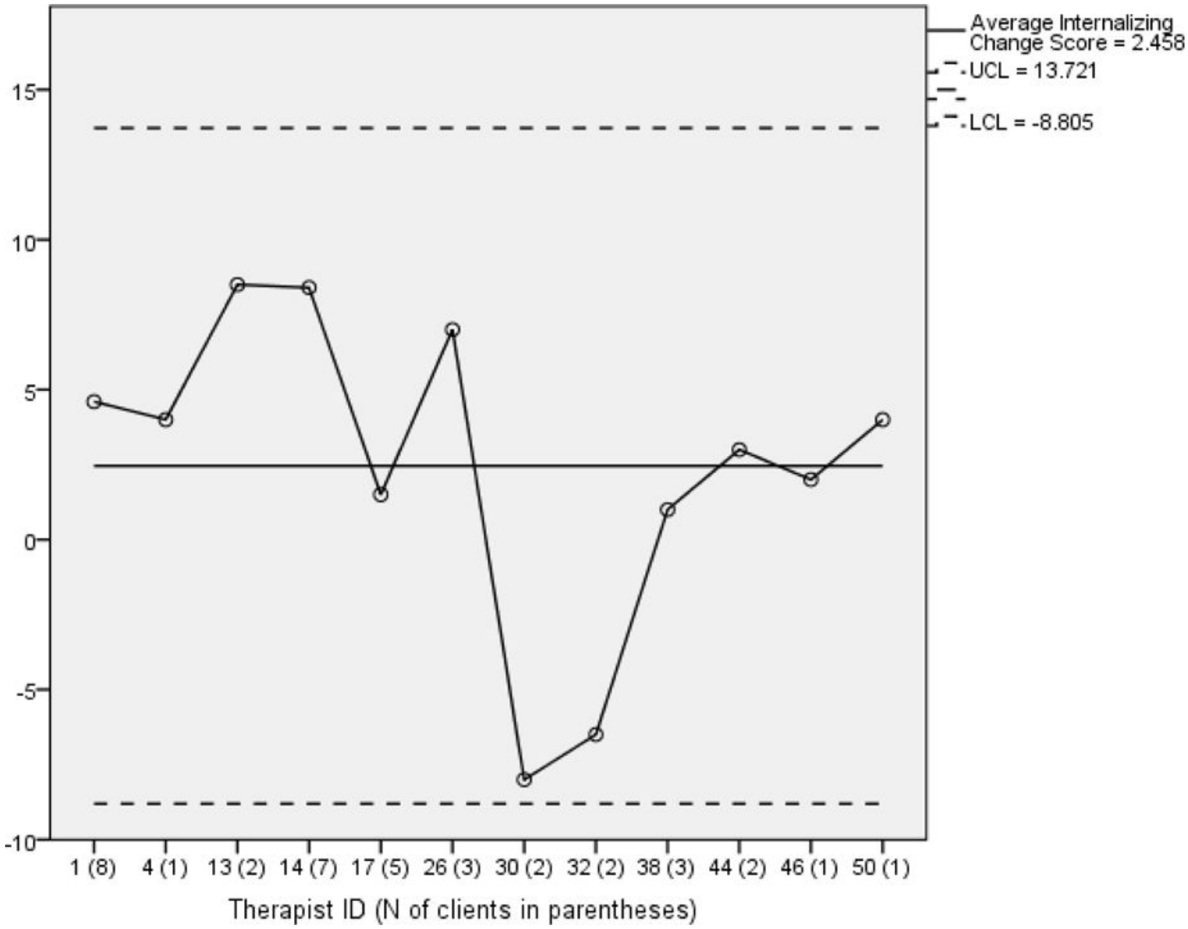

Figure 2.

SPC mean chart for change in Internalizing symptoms: Therapist variability in client outcomes.

Note. The numbers on the x-axis correspond to the Therapist identification code for each of the 12 therapists. Points on the chart are averages of Internalizing symptom change scores across clients for each therapist. UCL = Upper Control Limit for change in Internalizing symptoms derived from the data; LCL = Lower Control Limit for change in Internalizing symptoms derived from the data.

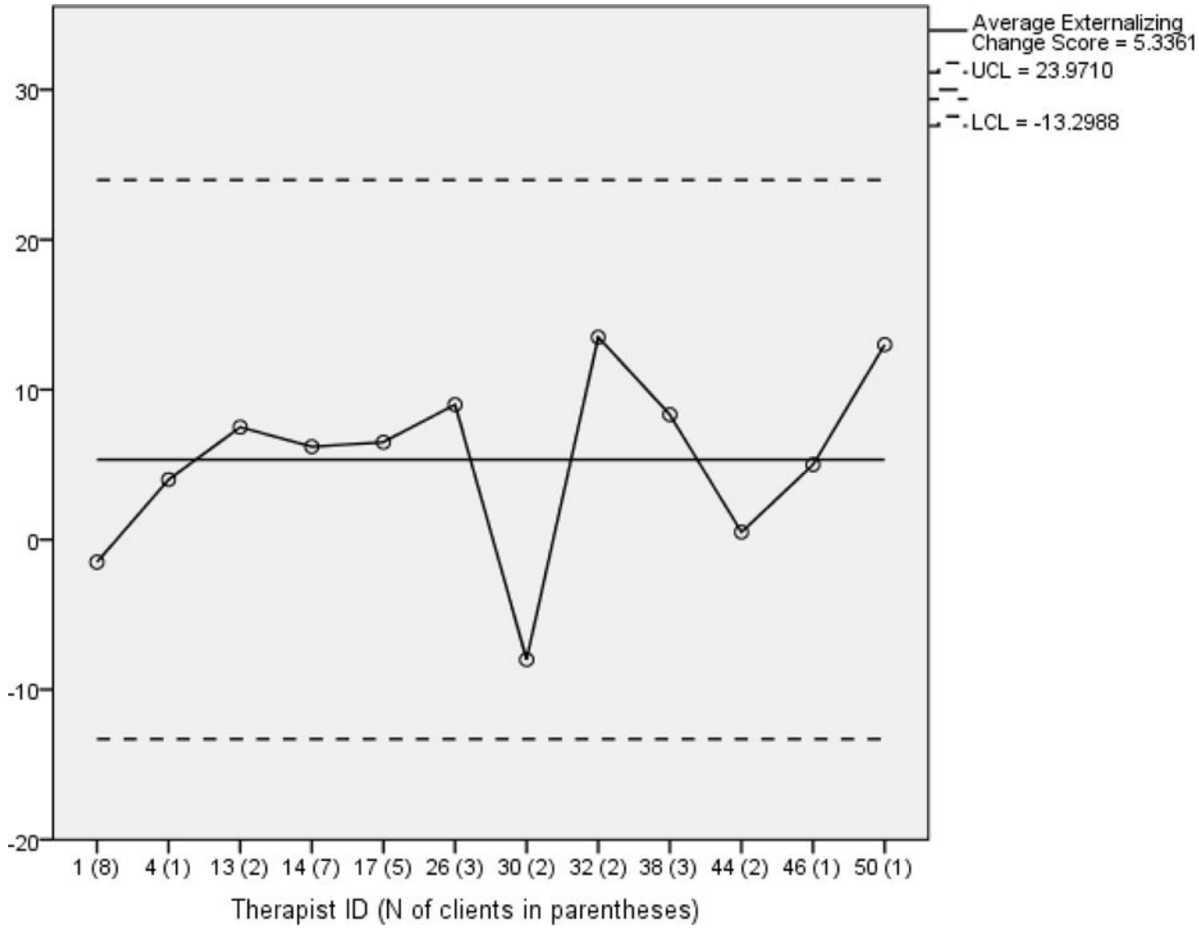

Figure 3.

SPC mean chart for change in Externalizing symptoms: Therapist variability in client outcomes.

Note. The numbers on the x-axis correspond to the Therapist identification code for each of the 12 therapists. Points on the chart are averages of Externalizing symptom change scores across clients for each therapist. UCL = Upper Control Limit for change in Externalizing symptoms derived from the data; LCL = Lower Control Limit for change in Externalizing symptoms derived from the data.

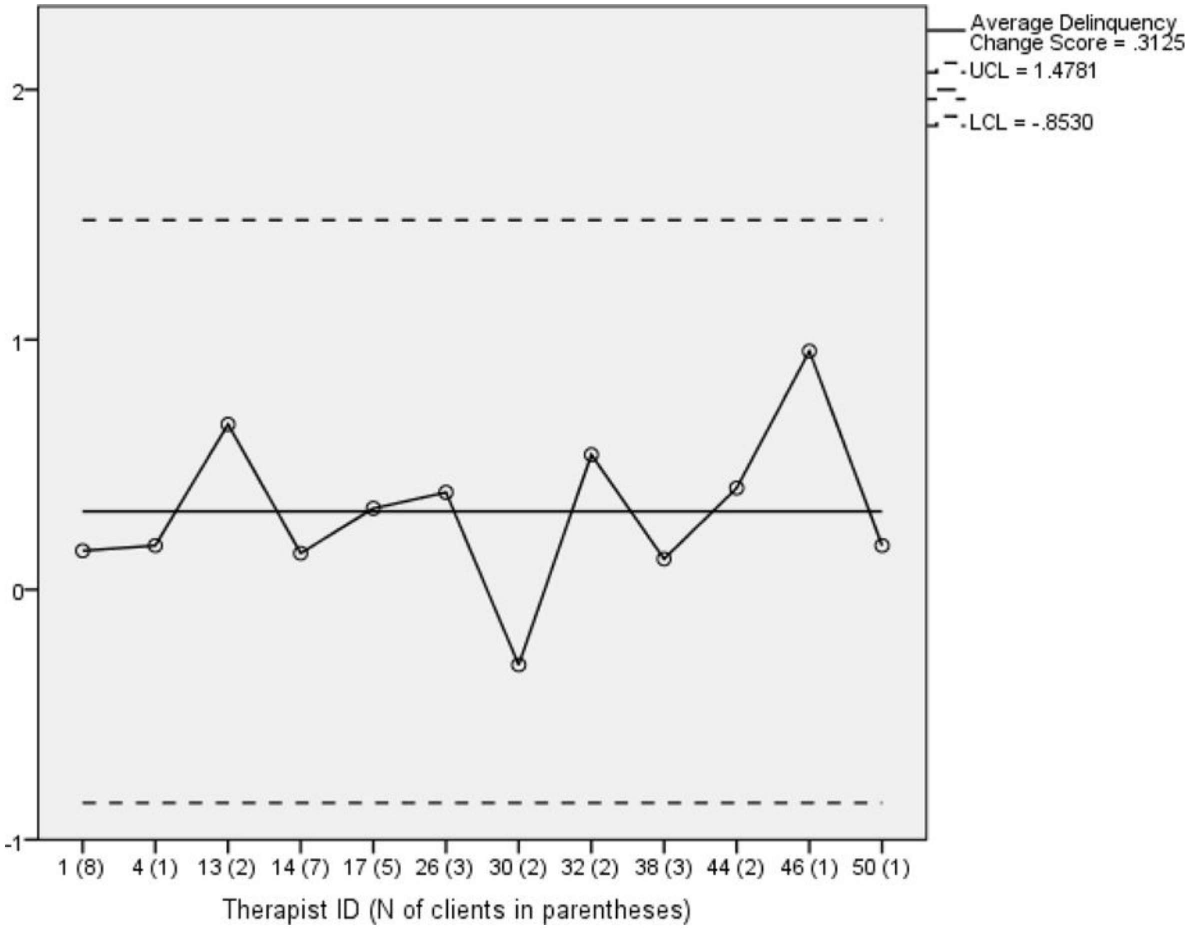

Figure 4.

SPC mean chart for change in Delinquency: Therapist variability in client outcomes.

Note. The numbers on the x-axis correspond to the Therapist identification code for each of the 12 therapists, with the number of clients treated by each therapist in parentheses. Points on the chart are averages of Delinquency change scores across clients for each therapist. UCL = Upper Control Limit for change in Delinquency derived from the data; LCL = Lower Control Limit for change in Delinquency derived from the data.