J Child Fam Stud
DOI 10.1007/s10826-014-9908-5

ORIGINAL PAPER

Psychometrics, Reliability, and Validity of a Wraparound Team Observation Measure

Eric J. Bruns · Ericka S. Weathers · Jesse C. Suter · Spencer Hensley · Michael D. Pullmann · April Sather

Abstract Wraparound is a widely-implemented team-based care coordination process for youth with serious emotional and behavioral needs. Wraparound has a positive evidence base; however, research has shown inconsistency in the quality of its implementation that can reduce its effectiveness. The current paper presents results of three studies used to examine psychometrics, reliability, and validity of a measure of wraparound fidelity as assessed during team meetings called the Team Observation Measure (TOM). Analysis of TOM results from 1,078 team observations across 59 sites found good overall internal consistency ($\alpha = 0.80$), but constrained variability, with the average team rated as having 78 % of indicators of model adherent wraparound present, 11 % absent, and 11 % not applicable. A study of $N = 23$ pairs of raters found a pooled Kappa statistic of 0.733, indicating substantial inter-rater reliability. Higher agreement was found between external evaluators than for pairs of raters that included an external evaluator and an internal rater (e.g., supervisor or coach). A validity study found no correlation between the TOM and an alternate fidelity instrument, the Wraparound Fidelity Index (WFI), at the team level. However, positive correlations between mean program-level TOM and WFI scores provide support for TOM validity as a summative assessment of site- or program-level fidelity. Implications for TOM users, measure refinement, and future research are discussed.

Keywords Wraparound · Children · Fidelity · Observation · Reliability

E. J. Bruns (✉) · E. S. Weathers · S. Hensley · M. D. Pullmann · A. Sather
Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, 2815 Eastlake Ave. E, Suite 200, Seattle, WA 98102, USA
e-mail: ebruns@u.washington.edu

J. C. Suter
Center on Disability and Community Inclusion, University of Vermont, Burlington, VT, USA

Introduction

Youths with serious emotional or behavioral disorders (SEBD) typically present with complex and multiple mental health diagnoses, academic challenges, and family stressors and risk factors (Cooper et al. 2008; Mitchell 2011; United States Public Health Service [USPHS] 1999). Such complex needs often garner attention from multiple public systems (e.g., child welfare, juvenile justice, mental health, education), each of which has its own mission, mandates, funding streams, service array, and eligibility requirements. Lack of coordination across these systems and fragmentation of service planning and delivery can exacerbate existing challenges for these youths, leading to costly and often unnecessary out of home placements. Thus, it has long been recommended that care coordination be provided that can integrate the multiple services and supports that a youth may receive across systems, as well as relevant services for caregivers and siblings (Cooper et al. 2008; Stroul 2002; Stroul and Friedman 1986; USPHS 1999).

Over the past two decades, the wraparound process has become the primary mechanism for providing care coordination to youths with SEBD and their families, and a central element in efforts to improve children's mental health service delivery for youths with complex needs (Bruns et al. 2010, 2013). Wraparound is a defined, team-based collaborative care model for youth with complex

Published online: 24 January 2014

behavioral health and other needs and their families. Driven by the preferences and past experiences of the youth and family, wraparound teams develop and oversee implementation of an individualized plan to meet their priority needs and goals. Wraparound teams consist of a heterogeneous array of professionals involved in service delivery, natural supports such as friends and extended family, and community supports such as mentors and members of the faith community (VanDenBerg et al. 2003).

At its most basic level, wraparound is defined by a set of ten principles (Family Voice and Choice, Team Based, Natural Supports, Collaborative, Community Based, Culturally Competent, Individualized, Strengths Based, Unconditional, and Outcome Based) and by a template for practice that consists of specified activities that take place over four phases of effort: engagement, planning, implementation, and transition (Bruns et al. 2010; Bruns et al. 2008, Walker et al. 2008b). There are approximately 800 unique wraparound programs in the United States serving approximately 100,000 children and families (Bruns et al. 2011). Wraparound has proven extremely popular among the children and families who are served in wraparound programs and initiatives, and although the majority of studies are quasi-experimental, has demonstrated positive results from 10 controlled, peer-reviewed studies (Bruns and Suter 2010; Bruns et al. 2010; Suter and Bruns 2009).

Fidelity Measurement in Wraparound

As for any innovative practice, effective implementation of wraparound requires attention to multiple levels of effort (Fixsen et al. 2005; Proctor et al. 2009; Walker et al. 2011). These include a hospitable community- and state-level policy and fiscal environment (Walker et al. 2008a, 2011), as well as a range of "implementation drivers," such as selection of appropriate staff, training, coaching, supervision, and quality and outcomes monitoring (Fixsen et al. 2005; Miles et al. 2011). Research on multiple behavioral health interventions (Glisson and Hemmelgarn 1998; Williams and Glisson 2013) as well as wraparound (Bruns et al. 2006) has demonstrated that without adequate attention to such community supports, organizational context, and implementation drivers, innovative practices are unlikely to succeed.

Fidelity monitoring—and effective use of results—is a fundamental implementation driver for behavioral health services (Fixsen et al. 2005; Schoenwald et al. 2011), including wraparound (Bruns et al. 2004). Fidelity is defined as the degree to which intervention delivery adheres to the original intervention protocol (Institute of Medicine 2001). Reliably and validly measuring adherence to fidelity is fundamental to understanding the results of research and evaluation studies, and can be used to support faithful implementation of effective treatments and services (Schoenwald 2011). With the increase in the emphasis of using research-informed practices to improve mental health outcomes, the use of fidelity measures to support these purposes has grown substantially (Schoenwald 2011).

Theory, research, and understanding of the fundamentals of effective fidelity measurement have also expanded greatly. For example, fidelity tools ideally provide a measure of program content (adherence to the components of the intervention) as well as process (e.g., skills associated with delivering the content and/or overall quality of service delivery) (Eames et al. 2008). In addition, attention to the reliability and validity of the measurement approach is important. For example, although provider self-monitoring and report can reduce data collection burden and help shape staff behavior, self-report tools are susceptible to response bias (Schoenwald et al. 2011; Eames et al. 2008). Similarly, collecting client or parent reports may be low burden and align with the goal of engaging families in the monitoring process; however, client/family reports can also be subjective and susceptible to response bias that limits variability.

Although wraparound was introduced to the field of children's mental health as a concept in the 1980s (Burchard et al. 2002; VanDenBerg et al. 2003), the availability of formal measures to assess adherence to wraparound fidelity was limited until the introduction of the Wraparound Observation Form (WOF; Nordness and Epstein 2003) and the Wraparound Fidelity Index (WFI) in the early 2000s (Bruns et al. 2004, 2005). The current version of the WFI (version 4.0) assesses fidelity via interviews with four types of informants—provider staff, parents/caregivers, youths, and team members. Based on responses, trained interviewers then assign ratings on a 0–2 scale on 40 items that assess content and process of the wraparound care coordination process (Bruns et al. 2009).

To date, the vast majority of wraparound fidelity monitoring has been conducted using the WFI. In addition to supporting system-, program, and practice-level implementation of wraparound, the WFI has also been used extensively in research. For example, the WFI successfully differentiated wraparound from case management in a recently completed randomized trial (Bruns et al. 2010; Bruns et al. 2014 in submission). Fidelity as assessed by the WFI has also demonstrated the association between model adherence and youth and family outcomes. For example, Bruns et al. (2005) found that overall fidelity was associated with improved behavior, functioning, restrictiveness of living, and caregiver satisfaction with services. Studies by Cox et al. (2010); Effland et al. (2011); and Vetter and Strech (2012) found positive associations

between adherence to wraparound principles as assessed by the WFI and improvements in child functioning.

Observation of Wraparound Teamwork

Although the WFI has been widely used by wraparound initiatives and researchers, as a self-report tool, it is susceptible to aforementioned limitations borne of subjectivity and response bias. Specifically, previous research has found that wraparound fidelity measures that rely on staff, parent, and youth report of implementation content and process can be susceptible to limited variation and "ceiling effects" that reduce utility of the tools and compromise psychometrics (Bruns et al. 2004; Pullmann et al. 2013). As an alternative or complement to interview or self-report, independent, direct observation methods can provide a rich account of behaviors and interactions of interest that reduces bias from treatment expectancy or overestimation (Aspland and Gardner 2003; Eames et al. 2008; Webster-Stratton and Hancock 1998).

The importance of facilitation skills and effective teamwork in wraparound care coordination makes observation and feedback of wraparound team meetings a highly relevant focus. This recognition yielded the development of and research on several wraparound observation measures by the late 1990s. Although not specifically designed to assess fidelity to the wraparound model, the Family Assessment and Planning Team Observation Form (FAPT) was designed to be used in multiple service delivery settings. The FAPT measures the family friendliness of the individualized treatment planning process for youth with severe emotional and behavioral problems and their families, and demonstrated good interrater reliability (Singh et al. 1997). The Wraparound Observation Form (WOF), adapted from the FAPT and designed to garner information on adherence to eight principles of wraparound as demonstrated in team meetings, was found to have good interrater reliability (Epstein et al. 1998; Nordness and Epstein 2003).

Although the FAPT and WOF were instrumental to promoting fidelity measurement for wraparound, neither were made widely available, norms or national means were never produced for use by the field as benchmarks, and both cited psychometric concerns such as ceiling effects (Epstein et al. 1998). Moreover, they were developed prior to the fully explicated version of the wraparound practice model that was produced in the mid-2000s by the National Wraparound Initiative to promote more consistent training, coaching, certification, fidelity monitoring, and practice (Walker and Bruns 2006; Walker et al. 2008a).

The Wraparound Team Observation Measure

To address these gaps, and to add to the range of implementation support tools and the research base on wraparound, the Team Observation Measure (TOM) was developed in 2006. To create the TOM, an initial item pool was developed by reviewing measures such as the FAPT and WOF and soliciting extensive expert input. A multi-round Delphi process (van Dijk 1990) was then undertaken with over 20 wraparound experts, who rated indicators for content and relevance (Bruns and Sather 2007), yielding an initial version with 78 indicators. An initial inter-rater reliability study of the TOM assessed agreement of pairs of observers for all TOM indicators for 15 wraparound team meetings. Results showed that the mean percent agreement between raters across all 78 indicators was 82 %. However, when employing a test statistic, Cohen's Kappa (Cohen 1960), that corrects for the likelihood of agreement by chance alone (high for TOM indicators given that there are only two response options), results showed a mean Kappa of only 0.464, indicating only moderate agreement between raters (Landis and Koch 1977). In 2009, these results were used to revise scoring rules to be more objective and clear and to eliminate indicators that were more difficult for the observers to score reliably. This revision resulted in the current version of the TOM with an updated scoring manual and 71 indicators organized into 20 items, each with 3–5 indicators.

To date, there has been only one published empirical study that used the TOM. Snyder et al. (2012) evaluated the implementation of Child and Family Team (CFT) meetings with families involved in the North Carolina child welfare system using the original version of the TOM. The study found higher TOM scores in counties that received resources to implement services in keeping with System of Care (SOC) principles (Stroul and Friedman 1986) providing evidence for validity of the TOM as well as evidence that SOC resources can have a positive impact on practice. The study also evaluated the internal consistency of the 20 TOM items, based on data collected in the study. Results found that despite having five or fewer indicators per item, nine out of 18 items (50 %) were found to have adequate internal consistency per the criterion of Cronbach's $\alpha > 0.60$. An unpublished study of the original TOM (Bruns et al. 2010) found significant site-level correlation between mean TOM Total Scores and mean WFI Total scores for eight sites in California that used both measures [$r(8) = 0.857$; $p < 0.01$], providing additional evidence for construct validity of the TOM.

The Current Study

To date, the TOM has been used in over 50 programs and wraparound initiatives in the U.S., Canada, and New Zealand. Despite substantial developmental work, one major empirically-based revision, and a national dataset of over 1,400 team meetings observed, the research base on the TOM remains limited. The children's services field

would benefit from a greater understanding of the psychometrics, reliability, and validity of the TOM. Moreover, because the TOM is used both by external evaluators as well as internal program staff (e.g., supervisors and coaches) it would be helpful to have a research-based understanding of differences in response patterns for different types of users, to guide decisions on how to use the measure. Finally, the research team would benefit from continued synthesis of information from a range of studies that can inform ongoing measure development and item and indicator refinement and/or elimination.

Toward these goals, the current paper aims to address five research questions, addressed through three separate substudies.

1. Based on analyses of our national TOM dataset, what is the variability and internal consistency of the TOM indicators and items? What are patterns of wraparound practice nationally as evaluated by TOM results from participating programs?
2. Based on results of two reliability studies conducted in different practice settings, what is the inter-rater reliability of the TOM? What are the differences in scoring patterns and inter-rater reliability for the two primary types of TOM users: external evaluators and internal evaluators?
3. Based on analyses of the association between TOM scores and WFI-4 scores in national sites that used both measures contemporaneously, what is the construct validity of the TOM?

Method

Measures

Team Observation Measure

The TOM (Bruns and Sather 2013) consists of 20 items, with two items dedicated to each of the 10 aforementioned Wraparound principles. Each item comprises 3–5 indicators of high-quality wraparound practice, each of which must be scored "Observed" or "Not Observed." (Not applicable is also an option for some indicators). A summary of the 20 items, how they relate to the 10 principles, and a sample indicator is presented in Table 1 (Bruns 2008).

Team Observation Measure observers are trained using the TOM Training Toolkit developed by the University of Washington Wraparound Evaluation and Research Team. Toolkit materials include a comprehensive manual and access to an online streaming video or DVD of a "mock" team meeting accompanied by a scoring key. Utilizing the

Table 1 Description of TOM items and sample indicators

Wraparound principle	Item	Sample indicator
Team Based	1. Team Membership and Attendance	"Parent/caregiver is a team member and present at the meeting"
	2. Effective Team Process	"Team meeting attendees are oriented to the wraparound process and understand the purpose of the meeting"
Collaborative	3. Facilitator Preparation	"The meeting follows an agenda or outline such that team members know the purpose of their activities at a given time"
	4. Effective Decision Making	"Team members reach shared agreement after having solicited information from several members or having generated several ideas"
Individualized	5. Creative Brainstorming and Options	"The team considers multiple options for tasks or action steps"
	6. Individualized process	"Team facilitates the creation of individualized supports or services to meet the unique needs of child and/or family"
Natural Supports	7. Natural and Community Supports	"Team members provide multiple opportunities for natural supports to participate in decision making"
	8. Natural Support Plans	"Brainstorming of options and strategies include strategies to be implemented by natural and community supports"
Persistence	9. Team Mission and Plans	"The team discusses or has produced a mission/vision statement"
	10. Shared Responsibility	"There is a clear understanding of who is responsible for action steps and follow up on strategies in the plan"
Cultural Competence	11. Facilitation Skills	"Facilitator reflects, summarizes, and makes process-oriented comments"
	12. Cultural and Linguistic Competence	"The team demonstrates a clear and strong sense of respect for the family's values, beliefs, and traditions"
Outcomes Based	13. Outcomes Based Process	"The team revises the plan if progress towards goals is not evident"
	14. Evaluating Progress and Success	"Objective or verifiable data is used as evidence of success, progress, or lack thereof"
Voice and Choice	15. Youth and Family Voice	"The team provides extra opportunity for the youth to speak and offer opinions, especially during decision making"
	16. Youth and Family Choice	"The family and youth have the highest priority in decision making"

Table 1 continued

Wraparound principle	Item	Sample indicator
Strengths Based	17. Focus on Strengths	"Team builds an understanding of how youth strengths contribute to the success of team mission or goals"
	18. Positive Team Culture	"The facilitator encourages ream culture by celebrating successes since the last meeting"
Community Based	19. Community Focus	"The team prioritizes access to services that are easily accessible to the youth and family"
	20. Least Restrictive Environment	"Serious challenges are discussed in terms of finding solutions, not placement in more restrictive residential or educational environments"

toolkit, trainees complete 4–5 training steps to ensure fluency and reliability. First, observers are trained on central concepts of wraparound that pertain to the scoring rules and target behaviors to be observed, as well as wraparound activities and phases (engagement and team preparation, initial plan development, plan implementation, and transition). Second, trainees review item-by-item scoring rules using the manual and/or an online training. Third, trainees are administered a knowledge test on the items, indicators, rules, and application of the rules. A score of 80 % is expected on the knowledge test before proceeding to practice observations.

Fourth, observers view the mock team meeting and score the TOM, comparing their scores and notes to the answer key provided. Eighty-five percent of items must be correctly scored on the "mock" team meeting before proceeding to in vivo practice observations. If a score of 85 % is not achieved, the trainee must observe two actual wraparound team meetings with an experienced peer or supervisor who has been trained to criteria before administering the TOM independently, comparing scores to the experienced peer or supervisor and reviewing the scoring rules when necessary to reach a final decision on scores on the 71 indicators. For trainees who score above 85 %, this step is recommended, but not required.

Wraparound Fidelity Index, Version 4.0

The WFI-4 was used in Study 3 as a means of testing the concurrent validity of the TOM. The WFI-4 is a 40-item administrated interview with parallel versions for wraparound facilitators, caregivers, youths, and other team members. All items are scored on a 0–2 scale (No, Sometimes/Somewhat, and Yes), and seven items are reverse-scored. Higher scores indicate increased wraparound fidelity. Four items are assigned to each of the ten principles of wraparound. The same items are also assigned to each of the four phases of wraparound implementation, from between six for the Engagement Phase to 15 for the Implementation Phase. Total measure scores are obtained through a sum of unweighted item scores. The measure is completed via 15- to 30-min semi-structured interviews, conducted by an interviewer who has been trained to criteria using a series of steps from a training toolkit (Sather and Bruns 2008) and an instrument manual (Bruns et al. 2009). Training includes administration guidelines, scoring keys, data entry instructions, and pre-recorded sample interview vignettes to assure interviewer's adherence to protocol. Interviewers must assign correct scores to 80 % or more items on three or more pre-recorded interviews.

Studies have found good psychometric properties for WFI scores (Bruns et al. 2005, 2010). Cronbach's alpha ranges from 0.83 to 0.92 for the four respondent types. A test–retest reliability study for the previous version of the WFI (WFI-3) found Pearson r correlations of 0.83 for facilitators and 0.88 for caregivers. A multiple regression indicated positive relationships between WFI scores at baseline and child behavioral strengths 6 months later as measured by the Behavioral and Emotional Rating Scale (Epstein and Sharma 1998), after controlling for baseline characteristics (Bruns et al. 2005). Another study (Effland et al. 2011) found positive relationships between WFI-4 scores and changes in scores on a standardized measure of functioning—the Child and Adolescent Needs and Strengths tool (Lyons et al. 2004). Intraclass Correlation for all three respondents has been found to be 0.51, indicating good inter-respondent agreement for a scale of this nature (Bruns et al. 2009).

Procedures and Results

Study 1: Wraparound Practice Examined Nationally as Assessed by the TOM

The purpose of the first study was to analyze TOM data from wraparound initiatives and programs across the US, in order to evaluate TOM psychometrics and item/indicator variability and better understand wraparound practice nationally.

Procedure

Between July 2009 and August 2012, 72 wraparound sites across the US participated in training and data collection using the updated (71 indicator) TOM. Thirteen sites submitted fewer than 5 TOM administrations. This raised

concerns that these sites may not be representative and/or have adequate experience with the measure, so these sites were removed from the dataset. Remaining sites were in eight states (CA, KY, MA, ME, NC, NJ, OH, PA) and run by local agencies (at the city or county levels) serving children and youth with emotional and behavioral challenges and their families. Nearly 90 % ($n = 53$) of these local sites were part of eight larger regional or statewide programs. The largest of these programs included 32 sites while the others ranged from 2 to 9 sites ($M = 7.88$, $SD = 10.05$). Most sites ($n = 40$) used internal observers (e.g., supervisors or coaches), while others ($n = 14$) used external observers (e.g., independent evaluators). A small number of sites ($n = 5$) used a mix of internal and external observers.

Once trained, observers used the TOM during wraparound meetings to measure whether each indicator was present, absent, or not applicable over the course of the entire team meeting. Data were submitted to the research team using the web-based *Wraparound Online Data Entry and Reporting System* (WONDERS), which also provides local users with an array of customized fidelity reports. WONDERS automatically scored each of the 20 items using a 5-point rating scale that corresponds to the number of indicators scored "Yes": 0 (*no indicators evident*), 1 (*fewer than half evident*), 2 (*half evident*), 3 (*more than half evident*), or 4 (*all indicators evident*). Indicators rated as *not applicable* were not included in the calculation of item scores.

All data submitted to the University of Washington (UW) Research Team were de-identified; local sites used ID numbers and maintained links to identifying information in a separate, secure location. Consent procedures varied across sites; however, the majority did not obtain formal approval from an Institutional Review Board (IRB) because the focus of data collection was quality improvement as opposed to formal research.

Participants

The 59 wraparound sites submitted TOM data for 1,366 wraparound team meetings. Most of these TOM administrations were for unique wraparound teams ($n = 1,269$); however, sites also submitted data on teams observed two times ($n = 86$), three times ($n = 9$), or four times ($n = 2$). We decided to use only the most recent TOM administration for each team, so no team was counted more than once in the dataset. In addition, we removed 191 TOM administrations for which data were missing for 20 % or more of the indicators.

These decisions resulted in a final sample of 1,078 teams observed using the TOM across 59 sites. Of the 746 teams that reported youth race and ethnicity, Wraparound teams supported youth identified as White (41 %; $n = 306$), Black (19 %; $n = 142$), Hispanic or Latino (19 %; $n = 141$), American Indian/Native American (10 %; $n = 75$), more than one race (7 %; $n = 52$), and "other" (4 %; $n = 30$). Of the 829 teams that reported youth gender, 66 % were identified as male ($n = 547$) and 34 % female ($n = 282$). Of the 747 teams that reported youth age, 12 % were 6 and younger ($n = 88$), 15 % were 7-9 years ($n = 114$), 22 % were 10-12 years ($n = 163$), 30 % were 13–15 years ($n = 220$), 19 % were 16–18 years ($n = 140$), and 3 % were 19 and older ($n = 22$).

Results

Team Characteristics

Teams were observed implementing different phases of the wraparound process: (a) engagement ($n = 98$), (b) planning ($n = 70$), (c) implementation ($n = 817$), and (d) transition ($n = 40$). Fifty-three teams did not report the type of team meeting. The finding that over 75 % of teams were in the implementation phase was consistent with recommended practice that engagement, planning, and transition phases should be relatively brief (approximately 2–3 weeks each) while the bulk of team work is accomplished during the ongoing meetings in the implementation phase. The number of team members present at the observed meeting ranged from 1 to 23 ($M = 6.08$, $SD = 2.24$, Median = 6.00). Nearly 90 % of teams had 8 or fewer members ($n = 940$). The most common team members present were parents or caregivers (92 %, $n = 989$), wraparound facilitators (90 %, $n = 971$), youth (69 %, $n = 742$), family advocates (56 %, $n = 599$), mental health providers (47 %, $n = 504$), other family members (28 %, $n = 304$), child welfare workers (25 %, $n = 266$), and school personnel (16 %, $n = 169$).

TOM Ratings

Team Observation Measure ratings for the 1,078 unique wraparound teams were very positive, indicating observers saw many examples of high fidelity wraparound at the participating sites. The average team was rated as having approximately 78 % of indicators of model adherent wraparound present, 11 % absent, and 11 % not applicable. In fact, only one indicator was rated as absent more often than present (*Natural supports are team members and present*). Five other indicators were rated as *not applicable* more often than present, also having to do with participation by natural supports (who were often not on teams) and discussions about residential placements (which frequently were not discussed). The most highly rated indicators were

from the items "youth and family voice and choice" (e.g., *Caregivers, parents and family members are afforded opportunities to speak in an open-ended way about current and past experiences and/or hopes about the future*) and "cultural and linguistic competence" (e.g., *Members of the team use language the family can understand*). Indicators least often observed reflected team membership (e.g., natural supports, school, or youth on the team) and items on outcomes based and "measuring progress" (e.g., *The team has set goals with objective measurement strategies*).

Average fidelity ratings were also high after aggregating indicators into TOM item and total scores. On the 0–4 scale, mean TOM item scores ranged from a low of 1.68 (*Natural and Community Supports, SD* = 1.76) to 3.89 (*Youth and Family Voice, SD* = 0.47), with a mean TOM Total Score of 3.45 (*SD* = 0.51).

Internal Consistency

The overall Cronbach's alpha for the TOM mean score was 0.80, considered a good to acceptably high rating of internal consistency (Bland & Altman, 1997). Examining individual items revealed that not all were equally contributing to a unified total score, with corrected item-total correlations ranging from 0.10 to 0.53 ($M = 0.38$, $SD = 0.11$). Four items had corrected item-total correlations below 0.30 (Items 1, 7, 15, and 20), and removing these items increased alpha to 0.82.

Team Observation Measure items comprise 3–5 indicators each, so we also examined corrected item-total correlations and Cronbach's alphas for each of the 19 items. Similar to Snyder et al. (2012) no alpha was calculated for the first item (*Team Membership and Attendance*), because the indicators ask if specific team members were present and were not expected to correlate meaningfully with each other. As shown in Table 2, alphas for TOM items ranged widely from 0.42 to 0.90 ($M = 0.59$, $SD = 0.14$). Average corrected item-total correlations for indicators within each item were similarly varied (ranging from 0.17 to 0.69, $M = 0.33$, $SD = 0.15$). Eight items had alphas greater than 0.60. Eight additional items could be improved somewhat by dropping single indicators, but still only 11 items had alphas greater than 0.60. Removing these eight indicators did not improve the alpha for the TOM mean score.

Study 2: Reliability of the TOM and Differences in Scoring Patterns for TOM Users

The purpose of this study was to assess the interrater reliability of the TOM and examine potential differences in scoring patterns between types of TOM observers; specifically, internal (e.g., supervisors and coaches) versus external (e.g., evaluators or managers) observers.

Procedure

Assessments of inter-rater reliability of the revised version of the TOM were completed using data obtained from N = 23 wraparound team meeting observations in Nevada and Washington. Twelve youth and families in Nevada and 11 youth and families in Washington were randomly selected for a team observation to be conducted by pairs of observers who had been trained to criteria on the TOM. The criterion for selection was having been enrolled in wraparound between two and 6 months. All observations in Nevada were conducted by two external observers: A research coordinator and a state-employed evaluator, between October 2009 and February 2010. Data in Washington were collected between April 2012 and August 2012 by eight observers: Four "external" observers were members of the UW research team; four "internal" observers were four wraparound coaches. Observations in Washington were conducted in pairs that consisted of one internal and one external observer.

All raters were trained to criteria as described above. All families were contacted by their wraparound facilitator and asked for permission prior to having their meeting observed. Upon the observers' arrival to the team meeting, families underwent informed consent. Study procedures were approved by Institutional Review Boards at the University of Washington and University of Nevada-Las Vegas.

Analyses

Cohen's Kappa was used to assess interrater reliability. Kappa measures the level of agreement between two raters compared to chance alone (Cohen 1960). Because of the small sample size and large number of TOM indicators, pooled Kappa was used to assess interrater reliability in the two sites. As suggested by De Vries et al. (2008), a pooled estimator of Kappa should be used when there are two independent observers, a small number of subjects, and a substantial number of measurements per subject. Pooled Kappa was calculated for TOM observations in Nevada, Washington, and data from the combined sites.

Results

TOM Reliability

Analysis of the 23 paired observations completed in Nevada and Washington suggest that the inter-rater reliability of the TOM improved as a result of the revision in 2009. Compared

Table 2 TOM Item means, standard deviations, and Cronbach's alphas

Wraparound Principle	TOM Item	Indic.	*n*	*M*	*SD*	α	α rev.
Team Based	1. Team Membership and Attendance	3	1,078	3.28	0.93	–	–
	2. Effective Team Process	4	1,078	3.71	0.62	0.42	–
Collaborative	3. Facilitator Preparation	4[a]	1,078	3.51	0.89	0.57	0.71
	4. Effective Decision Making	4[a]	1,078	3.66	0.69	0.47	0.59
Individualized	5. Creative Brainstorming and Options	3	1,052	3.30	1.33	0.82	–
	6. Individualized Process	4	1,078	3.70	0.64	0.43	–
Natural Supports	7. Natural and Community Supports	4	1,073	1.68	1.76	0.90	–
	8. Natural Support Plans	3	1,077	2.73	1.52	0.50	0.60
Persistence	9. Team Mission and Plans	4	1,078	3.68	0.67	0.44	–
	10. Shared Responsibility	3[a]	1,076	3.71	0.79	0.51	0.73
Cultural Competence	11. Facilitation Skills	4	1,078	3.62	0.83	0.62	–
	12. Cultural and Linguistic Competence	4[a]	1,077	3.85	0.48	0.48	0.50
Outcomes Based	13. Outcomes Based Process	3	1,034	3.14	1.44	0.76	–
	14. Evaluating Progress and Success	3	1,077	3.23	1.26	0.54	–
Family Voice & Choice	15. Youth and Family Voice	4[a]	1,078	3.89	0.47	0.64	0.66
	16. Youth and Family Choice	3[a]	1,066	3.74	0.73	0.48	0.56
Strengths Based	17. Focus on Strengths	4	1,078	3.48	1.02	0.75	–
	18. Positive Team Culture	4	1,078	3.68	0.75	0.59	–
Community Based	19. Community Focus	3	1,046	3.55	1.04	0.71	–
	20. Least Restrictive Environment	3[a]	779	3.88	0.59	0.63	0.70
	TOM Mean Score	71[b]	702	3.45	0.51	0.80	0.79

Indic. = number of indicators for each item; α = Cronbach's alpha; α rev. = Cronbach's alpha after one indicator removed

[a] One indicator was dropped from this item to increase item alpha

[b] Seven indicators were dropped to create revised TOM mean score

to the original 78-indicator version, which was found to have a pooled Kappa of 0.464, the pooled Kappa score for the revised TOM was 0.733, indicating substantial agreement, per conventions established by Landis and Koch 1977. Near-perfect agreement between paired raters was found for nearly all of the indicators making up the TOM items of Individualized Process, Natural and Community Supports, and Least Restrictive Environment. Indicators with substantial agreement were found for TOM items of Effective Decision Making, Team Mission and Plans, Shared Responsibility, Cultural and Linguistic Competence, Outcomes Based Process, Evaluating Progress and Success, Youth and Family Choice, and Focus on Strengths. Poor to fair agreement was found for at least one indicator for the TOM items of Effective Team Process, Facilitator Preparation, Creative Brainstorming and Options, Natural Support Plans, Facilitation Skills, Youth and Family Voice, Positive Team Culture, and Community Focus. Table 3 presents a summary of the level of agreement for all indicators in both substudies and overall.

Table 3 Differences in Level of Agreement for TOM Indicators

Level of Agreement	Nevada % of Indicators	Washington % of Indicators	Both sites % of Indicators
Almost perfect agreement	75	32	48
Substantial agreement	11	17	32
Moderate agreement	4	16	7
Fair agreement	3	11	7
Slight agreement	7	16	4
Poor agreement	0	8	1

Agreement criteria taken from Landis and Koch 1977

Differences by Rater Type

Differences in pooled Kappa scores were found between paired raters in Nevada and Washington. In Nevada, pooled Kappa for the 71 indicators was 0.843, indicating almost prefect agreement. However, in Washington, pooled Kappa was only 0.419, indicating moderate agreement between raters (Landis and Koch 1977). As shown in Table 3, differences in agreement were also found for individual indicators, with many more indicators found to be at substantial or near perfect levels of agreement for the Nevada sample that used two external raters (86 %) than for the Washington sample that paired an external with an

internal rater (49 %). These results suggest that raters with similar roles (e.g., external evaluators) may also be more likely to agree on TOM ratings than raters with different roles (e.g., a supervisor versus an external research team member).

To further test differences between rater types, we examined TOM scores by rater type for the Washington sample. Results found a mean TOM Total score of 3.43 ($SD = 0.34$) for internal raters versus 3.20 ($SD = 0.46$) for external raters, though due to small sample size this difference was not significant. Internal observers' ratings yielded higher item-level scores for 11 TOM items, compared to five for external evaluators (scores were equal for four items).

Study 3: Construct Validity of the TOM

The purpose of this study was to evaluate the association between TOM fidelity scores and fidelity as assessed by WFI-4 interviews, as an estimate of concurrent validity for the TOM.

Procedure

Team Observation Measure data were collected from July 2009 to August 2012 following the procedure outlined in Study 1. During this TOM data collection period, WFI-4 caregiver interviews were also conducted at 47 of the 59 wraparound sites that were using the TOM and included in Study 1. This smaller set of sites were from five states (CA, MA, NC, NJ, PA) and 44 of these sites participated as part of five larger wraparound programs or initiatives. Sites were required to use interviewers who had been trained to criteria using the *WFI-4 Interviewer Training Toolkit* (described above) prior to administering the WFI-4. Unlike the TOM, most sites used external WFI-4 interviewers ($n = 45$).

Participants

For the current study, only the WFI-4 caregiver total scores were used as they were the most readily available across these sites, and previous studies have shown the highest variability in caregiver WFI-4 scores compared to other respondents (Pullmann et al. 2013). For the 47 sites that had TOM and WFI-4 data, 918 teams had TOM administrations and WFI-4 interviews were conducted with caregivers on 1,098 teams. Unfortunately, many sites had separate tracking systems for TOM and WFI-4 administrations and did not use a common identification number for teams or families, restricting direct matching to just a small proportion of all cases. Only 138 teams had common ID numbers and both a TOM and WFI-4 caregiver interview conducted.

Results

As a follow-up to the first empirical comparison between TOM and WFI scores (Bruns et al. 2010), we conducted Pearson correlations between WFI-4 Caregiver and mean TOM scores at the program, site, and team (youth/family) levels. For program and site-level correlations, Pearson correlations were calculated between mean WFI-4 and TOM scores for the program or site. The correlation coefficient for mean TOM and WFI-4 caregiver scores at the program level was large and positive, but it was not significant (likely due to small sample size), $r(5) = 0.77$, $p = 0.12$. At the site level, the correlation remained positive but was smaller and again not significant, $r\ (47) = 0.20$, $p = 0.19$. Focusing on only the 138 teams that could be matched on TOM and WFI-4 administrations at the team level, the correlation was even lower, indicating no relationship between these two scores with this sample, $r\ (138) = -0.02$, $p = 0.79$.

Discussion

In this series of three studies, we sought to extend our understanding of response patterns, reliability, and validity for the Team Observation Measure, a measure of adherence to principles and prescribed activities of wraparound teamwork as observed in team meetings. Results of these studies, combined with previous research, suggest that the TOM as revised in 2009 has considerable psychometric strengths. Internal consistency was strong overall (Cronbach's $\alpha = 0.80$), and 16 of the 20 items had significant item-total agreement. Inter-rater agreement for the 71 indicators as assessed by pooled Kappa was 0.733, indicating substantial agreement, and was found to be meaningfully higher for the revised version of the TOM than its original 78-indicator version. Moreover, although sample size was small ($n = 12$), agreement when both raters were in external observer roles was near perfect. Finally, consistent with previous research, program-level mean Total TOM scores correlated very highly with mean Total WFI scores for the same programs, providing support for validity as a summative assessment of site- or program-level fidelity. Previous studies (Snyder et al. 2012) found that program (county) level TOM Total scores predicted level of investment in systems of care development, providing further support for validity.

At the same time, this series of studies uncovered potential weaknesses in the measure that can be addressed

in a future revision. Using the "Yes—No—Not Applicable" scale, only 11 % of indicators were not observed to be present, with 78 % observed and 11 % not applicable. Such positive responses are encouraging in that they suggest wraparound programs are successfully achieving adherence to basic activities of the model. The same patterns, however, produce restricted variability and "ceiling effects" in item-level and Total TOM scores which may reduce usefulness of the TOM as a quality improvement measure.

One explanation for these high scores may be the large number of sites that use internal raters, such as supervisors of the staff facilitating team meetings. Indeed, in Study 2, we found lower inter-rater reliability when pairing "internal" raters with "external" raters—university or state level evaluators. We also found that scores assigned by internal raters were higher (though not significantly so). Debriefs with program staff in these studies suggested that supervisors and internal coaches may assign higher and/or less objective ratings because they know the family and the teamwork that has occurred to date, and may base ratings on past activities undertaken by the team, even if no such evidence was evident in the meeting being observed. Internal observers also have relationships with staff being observed and know more about their overall skills, leading to their willingness to "give them the benefit of the doubt."

A second potential weakness of the TOM's structure is that only about half of the measure's 20 items demonstrate adequate internal consistency. This is consistent with the findings of Snyder et al. (2012) and is likely due at least in part to the small number of indicators (3–5) per item. Regardless, low internal consistency limits the usefulness of these item structures in research and will demand that TOM users disaggregate certain items into their constituent indicators, as was done by Snyder and colleagues in their 2012 evaluation, and/or use the TOM Total score in research and evaluation.

Finally, the current study revealed potentially important patterns of TOM validity at different levels of aggregation. In study 3, we found a large correlation between TOM and WFI-4 results at a program level, $r(5) = 0.77$, but a lower correlation at the site level, $r(47) = 0.20$, and no association, $r(138) = -0.02$, at the family or team level. As described above, these results support the TOM's validity and use as an overall quality assurance tool at a site or program level, and also suggest that there is a fundamental latent variable related to a program's quality or fidelity of wraparound implementation that can be gleaned by multiple types of data collection. At a family or team level, however, observation of team meetings may provide a different lens on the wraparound process than, for example, interviews with family and team members, which can more readily evaluate implementation of the process overall, including the many wraparound activities that take place between team meetings. This lack of association at lower levels of aggregation also suggests that evaluation or research focusing on family, team, or practitioner levels of implementation may need to employ multiple methods to get a full picture of fidelity.

Limitations

Certain limitations of the current study and the TOM must be recognized. First, although data were compiled from over 1,000 team meetings in 59 sites across the country, the majority of observations were conducted in a small number of large statewide wraparound initiatives with an investment in fidelity data collection and use. Thus, user sites are self-selected and may not be representative of wraparound implementing programs nationally. Second, even starting with a large national sample of over 1,000 team meetings, only 138 ultimately were able to be included in the correlational validity study. Thus, this relatively small sample may not be representative of all sites and programs using the TOM, possibly influencing results. Similarly, the inter-rater reliability study had a very small sample size ($n = 23$) and was conducted in only two sites, restricting our ability to determine significance of between-group differences. Finally, the TOM itself has a potentially major limitation as an instrument in that it primarily measures adherence to prescribed activities, not competence or overall quality, as is often recommended (Chambless and Hollon 1998; Nordness and Epstein 2003). This limited focus of the TOM may have influenced its psychometrics and findings from validity tests.

Implications

Results of this study extend those of previous research (Nordness and Epstein 2003; Singh et al. 1997) and suggest that wraparound fidelity as assessed by observation of team meetings can be conducted reliably. The study also reinforces previous research (Bruns et al. 2010; Snyder et al. 2012) that indicates overall scores from the TOM associate meaningfully with other criteria. Our team's experience working with collaborating user sites also suggests that the 71 TOM indicators provide managers, supervisors, and practitioners with useful and adequately detailed feedback on areas of needed improvement, additional resources, policy changes, and training.

At the same time, caveats have emerged from this and other research studies that may influence how the TOM is used. First, low internal consistency was found for many of the TOM's 20 items. While some of the 20 TOM items demonstrated adequate internal consistency, in general, the TOM item structure may primarily be used as a way of organizing the TOM and its results. Thus, while TOM

Total scores reliably tap into overall fidelity, finer grained analysis of TOM results may require examination of indicator-level data.

Second, the current study indicates that TOM data correlates negligibly with other sources of fidelity information at a family or team level. Thus, while the TOM may provide a valid overall portrait of wraparound implementation fidelity, measures that provide additional perspectives may be necessary to get a full picture of implementation for an individual youth or family, and possibly to get such information at small levels of aggregation such as practitioners or subsites within larger programs.

Finally, results of this series of studies suggest that TOM results may be influenced by the type of observer. Specifically, internal staff may be less reliable observers and may inflate scores by considering additional information. Administrators, evaluators and others designing quality assurance plans should consider this information as they weigh options for conducting observations. While using internal staff such as supervisors or coaches may increase the likelihood that the information will be directly applied to staff skill development and quality improvement (and possibly be more cost-effective), it may come at the expense of reliability.

Future Research

In addition to providing insights on how the current TOM might best be used, it also points to areas for continued research. For example, although evidence for concurrent and construct validity has now been produced, association with child and family outcomes has not yet been attempted, as has been done for the WFI (Bruns et al. 2005; Cox et al. 2010; Effland et al. 2011). Given the small sample sizes, replication of the findings showing differences in results for different types of raters will be important to attempt in the near future.

Most immediately, results can now be applied to a second revision of the TOM. Indicators may be deleted or revised to create a briefer, more reliable, and more useful version of the TOM. Ideally, indicators for a newly revised TOM will also be better organized into empirically-informed domains, all of which demonstrate adequate internal consistency and can be used to summarize results for quality improvement and research. Finally, it may be important to include indicators of practitioner competence and implementation quality, as is recommended for fidelity instruments. Improving our ability to implement wraparound effectively will be the highest overarching priority, as the approach continues to be a cornerstone of state and national efforts to improve care for children and youth with the most complex behavioral health needs.

Acknowledgments This study was supported in part by grant R34 MH072759 from the National Institute of Mental Health. We would like to thank our national collaborators and the dozens of trained TOM observers in these wraparound initiatives nationally. Thanks also to the wraparound initiatives in Clark County, Nevada and King County, Washington for their collaboration on the inter-rater reliability studies.

References

Aspland, H., & Gardner, F. (2003). Observational measures of parent child interaction. *Child and Adolescent Mental Health, 8*, 136–144.

Bruns, E. J. (2008). Measuring wraparound fidelity. In E. J. Bruns & J. S. Walker (Eds.), *The resource guide to wraparound*. Portland, OR: National Wraparound Initiative, Research and Training Center for Family Support and Children's Mental Health.

Bruns, E. J., Burchard, J., Suter, J., & Force, M. D. (2005). Measuring fidelity within community treatments for children and families. In M. Epstein, A. Duchnowski, & K. Kutash (Eds.), *Outcomes for children and youth with emotional and behavioral disorders and their families* (Vol. 2). Austin, TX: Pro-ED.

Bruns, E. J., Burchard, J. D., Suter, J. C., Leverentz-Brady, K. M., & Force, M. M. (2004). Assessing fidelity to a community-based treatment for youth: The wraparound fidelity index. *Journal of Emotional & Behavioral Disorders, 12*(2), 79.

Bruns, E. J., Pullmann, M. P., Brinson, R. D., Sather, A., & Ramey, M. (2014). *Effectiveness of wraparound vs. case management for children and adolescents: Results of a randomized study*. Manuscript submitted for publication.

Bruns, E. J., & Sather, A. (2007). *User's manual to the wraparound team observation measure*. Seattle, WA: University of Washington, Wraparound Evaluation and Research Team, Division of Public Behavioral Health and Justice Policy.

Bruns, E. J., & Sather, A. (2013). *Team Oversation Measure (TOM) manual for use and scoring. Wraparound fidelity assessment system (WFAS)*. Seattle, WA: University of Washington School of Medicine, Division of Public Behavioral Health and Justice Policy.

Bruns, E. J., Sather, A., & Pullmann, M. D. (2010). *The wraparound fidelity assessment system-psychometric analyses to support refinement of the wraparound fidelity index and team observation measure*. Paper presented at the the 23rd Annual Children's Mental Health Research and Policy Conference, Tampa, FL.

Bruns, E. J., Sather, A., Pullmann, M. D., & Stambaugh, L. F. (2011). National trends in implementing wraparound: results from the state wraparound survey. *Journal of Child and Family Studies, 20*(6), 726–735. doi:10.1007/s10826-011-9535-3.

Bruns, E. J., & Suter, J. C. (2010). Summary of the wraparound evidence base. In E. J. Bruns & J. S. Walker (Eds.), *The resouce guide to wraparound*. National Wraparound Initiative: Portland, OR.

Bruns, E. J., Suter, J. C., Force, M. M., Sather, A., & Leverentz-Brady, K. M. (2009). *Wraparound Fidelity Index 4.0: Manual for training, administration, and scoring of the WFI 4.0*. Seattle: Division of Public Behavioral Health and Justice Policy, University of Washington.

Bruns, E. J., Suter, J. C., & Leverentz-Brady, K. M. (2006). Relations between program and system variables and fidelity to the wraparound process for children and families. *Psychiatric Services, 57*(11), 1586–1593. doi:10.1176/appi.ps.57.11.1586.

Bruns, E. J., Walker, J. S., Bernstein, A., Daleiden, E., Pullmann, M. D., & Chorpita, B. F. (2013). Family voice with informed choice: coordinating wraparound with research-based treatment

for children and adolescents. *Journal of Clinical Child & Adolescent Psychology. (Online First Publishing),*. doi:10.1080/15374416.2013.859081.

Bruns, E. J., Walker, J. S., & The National Wraparound Initiative Advisory Group. (2008). Ten principles of the wraparound process. In E. J. Bruns & J. S. Walker (Eds.), *Resource guide to wraparound*. Portland, OR: National Wraparound Initiative, Research and Training Center for Family Support and Children's Mental Health.

Bruns, E. J., Walker, J. S., Zabel, M., Estep, K., Matarese, M., Harburger, D., et al. (2010b). The wraparound process as a model for intervening with youth with complex needs and their families. *American Journal of Community Psychology, 46*(3–4), 314–331. doi:10.1007/s10464-010-9346-5.

Burchard, J. D., Bruns, E. J., & Burchard, S. N. (2002). The wraparound process. In B. J. Burns, K. Hoagwood, & M. English (Eds.), *Community-based interventions for youth* (pp. 69–90). New York: Oxford University Press.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*(1), 7–18.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46.

Cooper, J. L., Aratani, Y., Knitzer, J., Douglas-Hall, A., Masi, R., Banghart, P., et al. (2008). *Unclaimed children revisited: The status of children's mental health policy in the United States*. New York: National Center for Children. in Poverty.

Cox, K., Baker, D., & Wong, M. A. (2010). Wraparound retrospective: Factors predicting positive outcomes. *Journal of Emotional & Behavioral Disorders, 18*(1), 3–13.

De Vries, H., Elliott, M. N., Kanouse, D. E., & Teleki, S. S. (2008). Using pooled kappa to summarize interrater agreement across many Items. *Field Methods, 20*(3), 272–282.

Eames, C. C., Daley, D. D., Hutchings, J. J., Hughes, J. C., Jones, K. K., Martin, P. P., et al. (2008). The leader observation tool: A process skills treatment fidelity measure for the incredible years parenting programme. *Child: Care, Health and Development, 34*(3), 391–400. doi:10.1111/j.1365-2214.2008.00828.x.

Effland, V. S., Walton, B. A., & McIntyre, J. S. (2011). Connecting the dots: Stages of implementation, wraparound fidelity and youth outcomes. *Journal of Child and Family Studies, 20*(6), 736–746. doi:10.1007/s10826-011-9541-5.

Epstein, M. H., Jayanthi, M., McKelvey, J., Frankenberry, E., Hary, R., Potter, K., et al. (1998). Reliability of the wraparound observation form: An instrument to measure the wraparound process. *Journal of Child and Family Studies, 7*, 161–170.

Epstein, M. H., & Sharma, J. M. (1998). *Behavioral and emotional rating scale: A strength-based approach to assessment*. Austin, TX: PRO-ED.

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network.

Glisson, C., & Hemmelgarn, A. (1998). The effects of organizational climate and interorganizational coordination on the quality and outcomes of children's service systems. *Child Abuse and Neglect, 22*(5), 401–421. doi:10.1016/S0145-2134(98)00005-2.

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academy Press.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lyons, J. S., Weiner, D. A., & Lyons, M. B. (2004). Measurement as communication in outcomes management: The child and adolescent needs and strengths (CANS). *The Use of Psychological Testing for Treatment Planning and Outcomes Assessment, 3*, 461–476.

Miles, P., Brown, N., & The National Wraparound Initiative Implementation Work Group. (2011). *The wraparound implementation guide: A handbook for administrators and managers*. Portland, OR: National Wraparound Initative.

Mitchell, P. F. (2011). Evidence-based practice in real-world services for young people with complex needs: New opportunities suggested by recent implementation science. *Children and Youth Services Review, 33*, 207–216. doi:10.1016/j.childyouth.2010.10.003.

Nordness, P. D., & Epstein, M. H. (2003). Reliability of the wraparound observation form-second version: An instrument designed to assess the fidelity of the Wraparound approach. *Mental Health Services Research, 5*, 89–96.

Proctor, E. K., Landsverk, J., Aarons, G. A., Chambers, D., Glisson, C., & Mittman, B. (2009). Implementation research in mental health services: An emerging science with conceptual, methodological, and training challenges. *Administration and Policy In Mental Health, 36*(1), 24–34.

Pullmann, M., Bruns, E. J., & Sather, A. (2013). Evaluating fidelity to the wraparound service model for youth: Application of item response theory to the wraparound fidelity index. *Psychological Assessment, 25*(2), 583–598.

Sather, A., & Bruns, E. J. (2008). *Wraparound fidelity index 4.0: Interviewer training toolkit*. Seattle: Division of Public Behavioral Health and Justice Policy, University of Washington.

Schoenwald, S. K. (2011). It's a bird, it's a plane, it's … fidelity measurement in the real world. *American Journal of Orthopsychiatry, 76*, 304–311.

Schoenwald, S. K., Garland, A. F., Chapman, J. E., Frazier, S. L., Sheidow, A. J., & Southam-Gerow, M. A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 32–43.

Singh, N. N., Curtis, W. J., Wechsler, H. A., Ellis, C. R., & Cohen, R. (1997). Family friendliness of community-based services for children and adolescents with emotional and behavioral disorders and their families: An observational study. *Journal of Emotional and Behavioral Disorders, 5*(2), 82–92. doi:10.1177/106342669700500203.

Snyder, E. H., Lawrence, N., & Dodge, K. A. (2012). The impact of system of care support in adherence to wraparound principles in child and family teams in child welfare in North Carolina. *Children and Youth Services Review, 34*(4), 639–647.

Stroul, B. A. (2002). *Issue brief—system of care: A framework for system reform in children's mental health*. Washington, DC: Georgetown University Child Development Center, National Technical Assistance Center for Children's Mental Health.

Stroul, B. A., & Friedman, R. M. (1986). *A system of care for severely emotionally disturbed children and youth*. Tampa, FL: University of South Florida, Tampa Research Training Center for Improved Services for Seriously Emotionally Disturbed Children and Georgetown Univ. Child Development Center, Washington D. C. Cassp Technical Assistance Center.

Suter, J. C., & Bruns, E. J. (2009). Effectiveness of the wraparound process for children with emotional and behavioral disorders: A meta-analysis. *Clincial Child and Family Psychology Review, 12*(4), 336–351.

United States Public Health Service (USPHS). (1999). *Mental health: A report of the Surgeon General*. Rockville, MD: U.S. Department of Health and Human Services Administration, Center for Mental Health Services, National Institutes of Health, National Institute of Mental Health.

van Dijk, J. (1990). Delphi method as a learning instrument: Bank employees discussing an automation project. *Technological Forecasting and Social Change, 37*, 399–407.

VanDenBerg, J., Bruns, E., & Burchard, J. (2003). History of the wraparound process. *Focal Point: A National Bulletin on Family Support and Children's Mental Health: Quality and Fidelity in Wraparound, 17*(2), 4–7.

Vetter, J. & Strech, G (2012, March). *Using the ohio scales for assessment and outcome measurement in a statewide system of care*. Paper presented at the 25th Annual Children's Mental Health Research and Policy Conference, Tampa, FL.

Walker, J. S., & Bruns, E. J. (2006). Building on practice-based evidence: Using expert perspectives to define the wraparound process. *Psychiatric Services, 57*, 1579–1585.

Walker, J. S., Bruns, E. J., Conlan, L., & LaForce, C. (2011). The National Wraparound Initiative: A community-of-practice approach to building knowledge in the field of children's mental health. *Best Practices in Mental Health, 7*(1), 26–46.

Walker, J. S., Bruns, E. J., & Penn, M. (2008a). Individualized services in systems of care: The wraparound process. In B. A. Stroul & G. M. Blau (Eds.), *The system of care handbook: Transforming mental health services for children, youth, and families*. Baltimore, MD: Paul H. Brookes Publishing Company.

Walker, J. S., Bruns, E. J., & The National Wraparound Initiative Advisory Group. (2008b). Phases and activities of the wraparound process. In E. J. Bruns & J. S. Walker (Eds.), *Resource guide to wraparound*. Portland, OR: National Wraparound Initiative, Research and Training Center for Family Support and Children's Mental Health.

Webster-Stratton, C., & Hancock, L. (1998). Training for parents of young children with conduct problems: Contents, methods and therapeutic processes. In G. E. Schaefer & J. M. Breismeister (Eds.), *Handbook of parent training* (pp. 98–152). New York: Wiley.

Williams, N. J., & Glisson, C. (2013). Testing a theory of organizational culture, climate and youth outcomes in child welfare systems: A united states national study. *Child Abuse & Neglect (Online first publication)*,. doi:10.1016/j.chiabu.2013.09.003.