Psychological Assessment
2013, Vol. 25, No. 2, 583–598

1040-3590/13/$12.00 DOI: 10.1037/a0031864

Evaluating Fidelity to the Wraparound Service Model for Youth: Application of Item Response Theory to the Wraparound Fidelity Index

Michael D. Pullmann, Eric J. Bruns, and April K. Sather
University of Washington

The wraparound process is a mechanism for multisystem planning and care coordination for youth with serious emotional and behavioral problems. Fidelity monitoring is critical to effective implementation of evidence-based practices in children's mental health, as it helps ensure that complex interventions like wraparound are implemented as intended. The 40-item Wraparound Fidelity Index, Version 4 (WFI-4; Bruns, Burchard, Suter, Leverentz-Brady, & Force, 2004) is the most frequently used measure of fidelity to the wraparound process, but analysis of its psychometric properties is insufficient. An item response theory approach, Rasch partial credit models for ordered polytomous data, was used on ratings from 1,234 facilitators, 1,006 caregivers, and 221 team members, focused on 1,478 youths (55% male). Results indicated the WFI-4 measured a unidimensional construct, with little evidence of item bias and good item and model fit. However, the item information curve was skewed, with most people endorsing high-fidelity responses, and several items had duplicative location estimates. A reduced 20-item measure is proposed. Internal reliability estimates for scores from this reduced measure were approximately equivalent to the longer measure. However, both versions would benefit from additional items located in the highest fidelity area of either version of the scale where scores by greater than half of our sample fall, but only 3 items are located.

Keywords: wraparound, item response theory, measurement, fidelity, system of care

It is estimated that between 10% and 20% of children and adolescents (approximately 15 million in the United States) experience a diagnosable mental health disorder, with six to eight million experiencing serious emotional disturbance (Friedman, Katz-Leavy, Manderscheid, & Sondheimer, 1998; Kazak et al., 2010). Only 20%–30% of youth who would benefit from services actually receive help (Kataoka, Zhang, & Wells, 2002), with much of what is provided not based on evidence for effectiveness (Hoagwood, Burns, Kiser, Ringeisen, & Schoenwald, 2001; Weisz, Jensen-Doss, & Hawley, 2006). Treatments that are provided are widely critiqued for not being culturally responsive, coordinated across systems, or adequately holistic to attend to the ecological context in which youth with complex mental health needs and their families need support (Bruns et al., in press; Farmer & Farmer, 2001; Tolan & Dodge, 2005). To address these issues, experts have called for improvements such as making empirically supported treatments more available, ensuring that services and supports are culturally responsive and well-coordinated, and engaging families more fully in the treatment process (Kazak et al., 2010; New Freedom Commission on Mental Health, 2003).

Over the past 20 years, the *wraparound process* has become a frequently applied mechanism for individualizing and coordinating the services and supports provided to youth with serious emotional and behavior problems (Walker, Bruns, & Penn, 2008; Weisz, Sandler, et al., 2006). Wraparound is an individualized, team-based, service planning and care coordination process. Wraparound team membership is individualized for each youth and family; teams usually consist of a wraparound facilitator, the caregiver (and child or youth, if they are mature enough to participate), and other pertinent individuals such as therapists, probation officers, teachers, friends, clergy, or others. The facilitator is trained to coordinate the process for the family, ensures meetings are adherent to the principals and activities of wraparound, acts as a consultant and advocate to the youth and family, and finds services in the community. The facilitator role is sometimes filled by the caregiver or a professional such as the therapist or caseworker. Caregivers and youth provide information about their needs, goals, and progress. Other team members provide information and support respective to their role in the family's life. The process aims to improve outcomes for youth with complex mental health problems through several mechanisms, including (a) integrating the efforts of the many systems and helpers who are involved, (b) basing the treatment plan on youth and family perspectives, (c) increasing social support received by the family, (d) actively integrating support for family and siblings in a holistic treatment plan, and (e) setting goals and monitoring progress (Bruns et al., in press; Walker et al., 2008).

This article was published Online First April 1, 2013.

Michael D. Pullmann, Eric J. Bruns, and April K. Sather, Department of Psychiatry and Behavioral Sciences, Division of Public Behavioral Health and Justice Policy, University of Washington.

This research was supported in part by National Institute of Mental Health Outcomes of the Wraparound Service Model Grant 1 R34 MH072759. The Wraparound Fidelity Index, Version 4, is a fidelity assessment instrument that is currently being programed into a web-based management information system for which Eric J. Bruns may receive future income. We acknowledge the local wraparound providers and sites who provided the data used in this study.

Correspondence concerning this article should be addressed to Michael D. Pullmann, Department of Psychiatry and Behavioral Sciences, University of Washington, Seattle, WA 98102. E-mail: pullmann@uw.edu

The evidence base for wraparound-type programs has expanded greatly in the past decade, with a recent meta-analysis (Suter & Bruns, 2009) finding significant, small to medium-sized effects in favor of wraparound across several outcomes domains, with particularly large effects (Cohen's d = .41–.55) found for outcomes related to maintenance of youth in community settings—a critical system- and family-level outcome. On the basis of such evidence—along with the model's popularity among family advocates and other stakeholders—wraparound is currently available in 90% of U.S. states (over half of which have statewide wraparound initiatives) to an estimated 100,000 youth (Bruns, Sather, Pullmann, & Stambaugh, 2011).

Fidelity Measurement and Wraparound Implementation

As evidence has mounted for the effectiveness of psychosocial interventions for youth, so has attention to *implementation fidelity*, defined as the degree to which the delivery of an intervention adheres to how the intervention was designed to be delivered (Dane & Schneider, 1998). Fidelity monitoring is critical to ensure a complex, multicomponent intervention is implemented as intended and to generalize effects found in research studies (Hogue, Liddle, Singer, & Leckrone, 2005). Fidelity measurement is also critical to research, including interpreting results of trials (i.e., Can we attribute study effects to the model?), conducting intervention process research (i.e., What about the model was responsible for effects?), and understanding practical questions about model application (e.g., What went on with this population that was different than with other populations?). Maintaining fidelity has been found to be critical to achieving positive outcomes across children's behavioral health, including prevention programs (cf. Domitrovich & Greenberg, 2000; Kam, Greenberg, & Walls, 2003), treatments for specific disorder types (cf. Alexander et al., 1998; Forgatch, Patterson, & DeGarmo, 2005), multimodal treatment models for youth with persistent and complex needs (cf. Alexander et al., 1998; Henggeler, Melton, Brondino, Scherer, & Hanley, 1997), and systems of care (Hernandez et al., 2001).

The Wraparound Fidelity Index

Wraparound implementation fidelity has been assessed using various methods, including team observation measures (Bruns & Sather, 2007; Epstein et al., 2003), document review (Bruns, Rast, Peterson, Walker, & Bosworth, 2006), and community stakeholder interviews (Walker & Sanders, 2011). The most commonly used measure is the Wraproud Fidelity Index (WFI; Bruns et al., 2004). The WFI was first developed in 1999 by John Burchard of the University of Vermont. The WFI has undergone three revisions, most recently in 2006, in order to assess adherence to the principles and core procedures of wraparound as specified by the National Wraparound Initiative (NWI). The NWI (www.nwi.pdx.edu) is a federally sponsored project that used national data and a *Decision Delphi*-facilitated expert consensus process to specify the procedures of the wraparound model (Walker & Bruns, 2006). The NWI established two primary mechanisms through which mechanisms of wraparound implementation can be communicated: (a) ten principles of wraparound service delivery and (b) four phases of wraparound implementation, including 32 core activities that are undertaken over the course of intervention with a youth and family.

In 2006, a 49-item version of the WFI was piloted in seven sites nationally that had used previous versions of the measure. All local evaluators reported it was an improvement over previous versions, and the pilot test found high reliability as measured by Cronbach's alpha for total score and adequate reliability for scores on the four phase subscales (Bruns, Suter, Rast, Walker, & Zabel, 2006). On the basis of this pilot study, problematic items that featured low variability, high ceiling effect, or difficulty in administration or interpretation were modified or removed. The resulting WFI, Version 4 (WFI-4) includes 40 items. See Appendix A for all measure items, categorized by the 10 principles and four phases. More detail on the measure is provided in the Method section.

The WFI-4 is now in use in over 50 sites across the United States, including several states that use the WFI-4 statewide. WFI data are used for different purposes by different provider agencies or wraparound initiatives. These uses range from research projects or program evaluation, where aggregate data from respondents are used to examine levels of fidelity across principles and implementation domains, to use in group or individual supervision and program assessment and development. WFI data are occasionally used at the individual level to examine the adherence to wraparound with staff, families, or teams as part of data-informed supervision. The current study was initiated because many staff at these local sites expressed a desire for a more parsimonious measure that could be administered more easily while still capturing essential implementation elements.

Item Response Models

Past psychometric analyses of previous versions of the WFI have used a classical test theory (CTT) approach. A contemporary alternative set of approaches collectively known as *item response theory* (IRT) has not been conducted with the WFI-4. In IRT models, respondents' estimated latent construct locations (e.g., the level of wraparound fidelity) depend on respondent scores *and* on item properties. Any individual's probability of rating an item in a particular way is a function of their location in the latent construct and the item's underlying relationship to the latent construct. In terms of wraparound fidelity, higher fidelity items are items that lower fidelity wraparound teams are unlikely to endorse.

Unlike CTT, IRT approaches allow the standard error of measurement to vary across items scores (Embretson & Reise, 2000). Hence, developers using IRT methods can calibrate tests so they focus more precisely on a particular range of the underlying construct, or more generally across the entire range of possible scores. This calibration process can help develop shorter measures with scores that have equal or better reliability estimates than longer forms, because multiple items measuring similar levels of a unidimensional construct are superfluous (de Ayala, 2009; Forkmann et al., 2009). Possible test bias is determined if particular items function differently within the model based on respondent type (Embretson & Reise, 2000).

IRT models that were inspired by the work of Georg Rasch (1980) differ in philosophy and approach to "traditional" IRT models inspired by the work of Lord and Novick (1968). They differ in the conceptualization and mathematical representation of the relationships among the latent construct, the model, and the

observed data, and some authors consider them to be different classes of approach (Andrich, 2004; Bond & Fox, 2007; Rasch, 1980). Mathematically, traditional one-parameter models do not differ from Rasch models when modeling the single parameter of item location, described above (de Ayala, 2009). Traditional models may also feature additional parameters. Because Rasch models do not feature these parameters, comparisons between persons are invariant over the specific items used, and comparisons between items are invariant over the specific data used.

Therefore, Rasch modeling facilitates the construction of comparable measured scores, whereas traditional IRT models facilitate accounting for the most variance in the data. Rasch models allow the tallying of a total score that completely characterizes person-level trait, unlike more complex traditional models. This is important because the goal for this project was to create an easy-to-score measure that wraparound facilitators, families, and others could calculate and understand in settings without access to a computer or scoring chart. More complex models would not have facilitated this goal.

Partial Credit Model and Assumptions

The partial credit model (PCM), which is the Rasch IRT model we used for the current analysis, incorporates graduated responses (i.e., "partial credit"); this makes it an appropriate model for ordinal response categories such as the WFI scale. The PCM estimates threshold parameters for each item, that is, levels at which respondents' predicted probability of responding changes from one category to the next (i.e., from "No" to "Sometimes/Somewhat" to "Yes"). The PCM makes no assumptions about the spread of the category threshold parameters; hence, the distance in estimated fidelity levels between categories does not have to be consistent from item to item.

Assumptions for the Rasch-based PCM include the following: The measure must have essential unidimensionality; items must have local independence, or that the items are independent after consideration of the respondent's location parameter (technically, this is subsumed under the unidimensionality assumption); items must equally discriminate for all respondents; and randomness is normally distributed as indicated by an approximation to a standardized residual mean of zero and a standard deviation of 1.0. Statistical assessment of unidimensionality is a matter of degree—no model fits data perfectly. As Embretson and Reise (2000) wrote, "It is arguable that purely unidimensional scales can only be created for the narrowest of constructs (e.g., self-esteem in eighth-grade math class)" (p. 262). Evidence indicates that parameter estimation is relatively unaffected by minor violations of unidimensionality (Childs & Oppler, 2000; de Ayala, 2009; Embretson & Reise, 2000; Harrison, 1986), especially in certain situations. Violations of unidimensionality are of little concern if there is only one dominant dimension with several minor dimensions, and minor dimensions are correlated at $r > .4$ (Kirisci, Hsu, & Yu, 2001). The WFI may seem like a multidimensional construct, as it is created on the basis of 10 principles and four phases. However, the WFI is not measuring the wraparound principles and process per se; rather, it measures the team's *adherence* to the well-defined wraparound process. This unidimensional concept of adherence is the "fidelity" that the WFI measures.

Rationale for Current Study

Despite its widespread use, no prior studies of any of the WFI versions have used an IRT model-based approach. The goal of the current study was to use a Rasch-based IRT method to provide an in-depth analysis of the WFI-4, provide recommendations to improve the functioning of the measure, and explore a shorter, easier-to-use version to ease response burden and facilitate use in the field and by laypersons. Though the WFI-4 is intended to evaluate the perspectives of several informants (i.e., the facilitator, caregiver, and other team members such as friends, therapists, and teachers), it is possible that these individuals may exhibit differential patterns of responding that may hinder comparability of forms across respondent type. We sought to improve the WFI-4 by removing items revealed to be irrelevant, biased, or superfluous.

Method

Measure

The WFI-4 is a 40-item administrated interview with separate, parallel versions for three types of respondent: wraparound facilitators; caregivers; and other team members who are not facilitators, caregivers, or youth (e.g., therapists, teachers, or friends of the family). Versions differ only in target wording (e.g., "your child" or "your family" compared with "the child" or "the family"). The WFI also has a version for youth that contains 32 items, including five items that have no direct corollary in the other three versions, and many of the "parallel" items use very different wording. Hence, for this study, we did not examine the youth version. All items are scored as Yes, Sometimes/Somewhat, and No (usually 2, 1, and 0, but seven items are reverse scored). Higher scores indicate increased wraparound fidelity. Development of the measure included four items assigned to each of the 10 principles of wraparound and six to 15 items assigned to each the four phases of wraparound implementation (1. Engagement and Team Preparation, 2. Initial Planning, 3. Implementation, and 4. Transition).

Total measure scores are obtained through a sum of unweighted item scores. Multivariate inferential analyses have generally used the total score for each respondent type, although some applications have combined respondent scores into an "average team score." Principle or phase scores have been calculated by averaging the items for each factor. The WFI-4 is available from the authors upon request, and Appendix A details the measure. Although the measure was developed with the intention that items adhered to the 10 principles, confirmatory factor analyses (CFAs) on caregiver, facilitator, and youth scores (separate or combined) from an earlier version of the WFI did not support a 10-element model due to extremely high intercorrelations among many of the principle factors (Suter, 2006). No clear factor structure emerged from these analyses using this earlier version, and the measure was substantially revised following these analyses.

The measure is completed via a 15- to 30-min semistructured interview with the caregiver, facilitator, or other team member, conducted by an interviewer who has been trained to interrater reliability criteria using a series of steps from a training toolkit (Sather & Bruns, 2008) and an instrument manual (Bruns, Suter, Force, Sather, & Leverentz-Brady, 2009). Training includes administration guidelines, scoring keys, data entry instructions, and

prerecorded sample interview vignettes to assure interviewer's adherence to protocol. Interviewers must assign correct scores to 80% or more items on three or more prerecorded interviews; however, overall interrater reliability studies have not been conducted.

Studies have found good psychometric properties of WFI scores for current and previous versions of the measure (Bruns, 2010; Bruns, Burchard, Suter, & Force, 2005). Scores reveal high internal consistency with WFI-4 Cronbach's alphas of .92 for caregivers, .83 for facilitators, and .86 for team members, and good test–retest reliability with WFI-3 Pearson *r* correlations of .83 for facilitators and .88 for caregivers. Concurrent validity estimates have been good, with a Pearson *r* of .86 when correlated at a mean site level with the Team Observation Measure, an observational tool for measuring fidelity to wraparound (Bruns & Sather, 2007; Bruns, Sather, & Pullmann, 2010; Sather, Cox, & Bruns, 2009). A multiple regression indicated positive relationships between WFI-2 scores at baseline and child behavioral strengths 6 months later, as measured by the Behavioral and Emotional Rating Scale (Epstein & Sharma, 1998), after controlling for baseline characteristics (Bruns, Suter, Force, & Burchard, 2005). Another study found positive relationships between WFI-4 scores and a decrease in youth needs (Effland, Walton, & McIntyre, in press), as measured by the Child and Adolescent Needs and Strengths measure (Lyons, Weiner, & Lyons, 2004). Studies have also found positive correlations between WFI scores and system factors theorized to be associated with successful wraparound implementation (Bruns, Suter, & Leverentz-Brady, 2006; Effland et al., in press). The intraclass correlation coefficient (ICC) for all three respondents for total score data from the current study is .51, indicating good interrespondent agreement for a scale of this nature. For instance, this ICC is stronger than cross-informant agreement found for major child behavior scales (Achenbach, McConaughy, & Howell, 1987).

Procedure

Data were collected by wraparound programs at sites throughout the United States in collaboration with the Wraparound Evaluation and Research Team (WERT; see http://depts.washington.edu/wrapeval) between February 2007 and June 2009. Interviews were conducted by local evaluators trained to interview protocol and interrater reliability criteria as described above. Participants may have been interviewed multiple times; for these analyses, we used the first interview. Ideally, the first interview is conducted between 2 and 6 months after a youth's enrollment in wraparound; however, procedures varied across different implementation sites, which highly skewed the distribution (range = 0–36 months, mode = 6, M = 11, SD = 7.4). Data were entered by local sites and submitted to WERT. Though this was a convenience sample (wraparound providers voluntarily chose to use the WFI-4 and participate in submitting scores to WERT), IRT approaches do not require randomly selected participants or a population- or time-based representative sample. The only sample-based requirement is that it be sufficiently heterogeneous to capture a range of scores (Embretson & Reise, 2000).

Participants

Data were submitted from 41 different wraparound sites, which were nested within 25 projects, including several statewide wraparound implementation initiatives. Data from 20 to 332 teams/youth came from each project, with an average of 59.1 teams from each project. Data were obtained from a total of 2,461 people, including 1,234 facilitators, 1,006 caregivers, and 221 team members. Frequencies of item responses by respondent type are located in Appendix B. A total of 1,478 unique wraparound teams/unique youths were assessed. The youths of focus were 55% male, 47% non-Hispanic White, 20% African American, 2% American Indian or Alaska native, and less than 1% Asian/Native Hawaiian/Pacific Islander. An additional 7% identified as mixed race, and 7% identified as another category. Fourteen percent were missing data on race. Two percent identified as Hispanic ethnicity (White, mixed race, or "other"). Youth age ranged from 1 to 21 years old, and the distribution was negatively skewed (M = 12.3, SD = 5; Mdn = 13.7; mode = 15.5). Teams ranged from two to 11 members, with an average of 4.8 (SD = 1.6), including facilitators (99%), birth mothers (58%), birth fathers (24%), foster parents (21%), youths (45%), siblings (17%), grandparents (20%), other family members (15%), therapists or other mental health workers (53%), family support partners or advocates (33%), teachers or school representatives (23%), friends (9%), child welfare caseworkers (22%), attorneys (14%), mentors (9%), probation officers (7%, and, at less than 5%, other roles such as medical providers, court-appointed special advocates, clergy, adoptive parents, and other natural supports.

Analytic Strategy

Analyses proceeded in a series of five stages. To determine essential unidimensionality, the first stage was to conduct an exploratory factor analysis (EFA) and a CFA for the overall sample and for important subgroups (caregiver, facilitator, and team member). Good fit for the EFA is determined by examining scree plots (described here as a first- to second-Eigenvalue ratio of at least 3:1), and matrix residuals with a mean near zero and standard deviation of .05 and below. Good fit for the CFA is determined by a root-mean-square error of approximation (RMSEA) less than .06, and a Tucker-Lewis Index (TLI) and comparative fit index (CFI) close to .95 (Hu & Bentler, 1999). We explored whether the subgroup factor loadings were similar using the congruence coefficient (Guadagnoli & Velicer, 1991; Lorenzo-Seva & Ten Berge, 2006), which is an index of the similarity between factors. Congruence coefficients above .95 are considered equal or virtually identical, and coefficients above .90 are considered to have a high degree of similarity.

The second stage of analysis was to build and examine the fit of the IRT model. Preliminary exploration of the data revealed that assuming a common set of thresholds for the item response set across all items was inappropriate, based on average item infit and outfit mean squares (de Ayala, 2009). Hence, we applied a Rasch PCM (Masters, 1982), which is appropriate for ordered polytomous data with varying item threshold sets. Estimates were obtained through joint maximum likelihood estimation using the WINSTEPS program, Version 3.69 (Linacre, 2009b).

Estimates of overall model fit to Rasch assumptions were examined, as well as overall model data fit for persons and overall

model data fit for items. Person- and item-mistfit analyses were conducted. Differential item functioning (DIF) was investigated in order to uncover items exhibiting statistical bias, or that performed differently on the basis of respondent type. DIF may reflect honest differences among respondent types, in which case the items may remain relevant to measuring the construct and might be retained, or it may reflect hidden bias. All measures rely on an assumption that respondent types do not differ in how they understand the item—if DIF indicates a violation of this assumption, then items detract from the functioning of the measure and should be removed (de Ayala, 2009; Embretson & Reise, 2000). As recommended by de Ayala (2009), the authors (including a nationally recognized wraparound expert) qualitatively interpreted any items with differential functioning to determine whether the DIF was relevant or irrelevant to the construct of wraparound fidelity. We jointly considered magnitude of DIF, item misfit, and possible reasons for DIF in order to identify items to remove from the measure.

In the third stage, we sought to shorten the measure. The model was rerun using the remaining items. We examined item locations, plotted a person-item fidelity-level map, and examined the total information curve. This exercise helped identify items measuring the same level of fidelity, which adds little in terms of measuring fidelity. Of this pool of duplicative items, we identified candidates for removal through a combination of factors, including item misfit scores, point-measure correlation estimates, item characteristic curves, and face validity. Because the removal of items may affect fit statistics, item removal proceeded in iterative blocks. Fourth, after these items were deleted, the PCM was rerun to obtain final estimates. Fifth, reliability estimates for scores from the initial and shortened measure were compared using Cronbach's alpha statistics.

Results

Factor Analyses

The EFAs were conducted on the individual scores for each respondent on each item using maximum likelihood estimation. Results indicated reasonable fit. Table 1 depicts the EFA fit statistics by group. The first- to second-Eigenvalue ratio for most groups was good, though for facilitators it was slightly small (2.8:1); mean residuals were good for the overall sample and all groups (close to 0), and residual standard deviations were slightly high for all groups (between .05 and .06), especially team members (.083). All goodness-of-fit tests were significant. Exploratory analyses, not depicted here for space reasons, indicated that over 90% of correlations among possible minor dimensions for two, three, four, and five possible factors were $r > .4$. These results indicate that the factor structure was accounted for by one dominant factor with several highly correlated minor factors.

The CFA was conducted on the individual scores for each respondent on each item using tetrachoric correlations for ordinal response categories, specifying a single factor with factor loadings permitted to vary. Two indices revealed reasonable model fit (RMSEA = .057 and TLI = .945), whereas one failed to exceed a critical value (CFI = .740). Closer inspection of the fit indices indicate that this may have been due to a few misfitting items. Fit for subgroup analyses were similar, with caregiver RMSEA = .063, TLI = .955, CFI = .847, facilitator RMSEA = .051, TLI = .874, CFI = .765, and team member RMSEA = .061, TLI = .896, and CFI = .827. Hence, for the overall sample and subgroups, unidimensional model fit was not strong, though at least one measure of fit exceeded critical values for all subgroups except team members, which had the smallest sample size. Specifying items with lower factor loadings into second or third factors did not improve fit for the overall sample or subgroups. These results are not inconsistent with the conclusion from the EFA that the factor structure is composed of one dominant factor and several highly correlated minor factors.

Table 2 depicts the EFA and CFA factor loadings for the overall sample (sorted in order of the location estimates from Table 3 to facilitate comparability). For the CFA on the overall sample, factor loadings for most items were moderate to high, ranging from .50 to .84, but five items were between .30 and .50, and two items were below .30. The CFA factor loadings for the subsamples were similarly moderate to high. However, for the EFA on the overall sample, factor loadings for most items were low to moderate, with six items below .30 (see Table 2). Factor loadings for caregivers were similar, but factor loadings for facilitators and team members were lower.

Most cross-informant factor loadings for the CFA were highly similar. Three items with an overall factor loading greater than .30 had at least one subgroup loading less than .30. However, there were several differences between cross-informant factor loadings in the EFA—16 items with an overall factor loading greater than .30 had at least one subgroup loading less than .30. Therefore, in order to test whether factor structures were similar, we calculated a series of pairwise congruence coefficients testing each subgroup set of CFA and EFA factor loadings against each other subgroup and the overall factor loadings (see Table 2). Most congruence coefficients were above .95, which indicates that the factor structures between overall score and caregiver, facilitator, and team

Table 1
Exploratory Factor Analysis Fit Estimates

Variable	1st EV	2nd EV	Ratio of 1st to 2nd EVs	Residual *M*	Residual *SD*	Goodness-of-fit test: χ^2 (*df*)
Overall	9.79	2.04	4.8:1	.003	.052	5,566 (740)*
Caregiver	10.89	2.11	5.2:1	.004	.057	2,734 (740)*
Facilitator	5.96	2.11	2.8:1	.001	.056	2,846 (740)*
Team member	7.30	2.38	3.1:1	.002	.083	1,180 (740)*

Note. EV = Eigenvalue.
* $p < .001$.

Table 2
Factor Analysis Loadings

Variable	EFA single-factor loadings				CFA single-factor loadings			
	Overall	CG	Fac.	TM	Overall	CG	Fac.	TM
Congruence coefficient								
Overall	—	.995	.952	.962	—	.996	.985	.981
Caregiver		—	.932	.943		—	.973	.971
Facilitator			—	.948			—	.976
2.3 Mostly professional services*	.116	.087	.195	.212	.193	.157	.249	.266
4.1 Develop transition plan	.300	.216	.319	.311	.396	.309	.381	.371
3.6 Friend advocate participation*	.199	.175	.289	.262	.298	.261	.385	.387
2.5 Strategies get child activities	.391	.447	.277	.269	.514	.568	.391	.453
3.3 Child involved activities	.399	.443	.317	.221	.529	.569	.429	.406
4.2 Develop youth friendships	.484	.452	.580	.545	.664	.626	.707	.722
1.4 Family select team members	.438	.325	.333	.296	.533	.411	.478	.435
4.3 Help child solve problems	.507	.479	.554	.454	.670	.634	.683	.616
3.4 Increase informal support	.572	.602	.472	.489	.684	.727	.579	.612
4.7 Family succeed without WA*	.146	.019	.422	.367	.298	.143	.529	.469
3.13 Team end before family ready*	.205	.155	.242	.333	.339	.285	.362	.431
1.5 Difficult to get team members attend*	.217	.299	.299	.076	.317	.412	.379	.125
4.4 Help child prepare for transitions	.554	.541	.451	.518	.708	.699	.629	.649
2.10 Family not highest priority in design	.443	.547	.272	.381	.552	.638	.413	.551
2.2 Written team vision	.422	.467	.298	.190	.534	.564	.447	.241
3.1 Decisions made family not there	.475	.567	.319	.160	.581	.667	.439	.260
3.9 Team assign review tasks	.561	.545	.424	.491	.677	.652	.562	.651
2.8 Crisis or safety plan	.465	.456	.272	.296	.622	.613	.452	.510
4.6 Family get supportive relationships	.578	.634	.582	.591	.724	.763	.742	.710
3.8 Services difficult to access*	.323	.342	.254	.155	.442	.454	.390	.230
2.9 Keep child in community	.349	.381	.374	.430	.506	.530	.537	.510
1.6 Identify crises or dangerous situations*	.509	.487	.340	.408	.664	.636	.546	.571
2.6 Members no role implementing*	.235	.238	.242	.284	.383	.406	.349	.447
4.5 Team restart if needed*	.356	.459	.290	.084	.542	.627	.472	.342
4.8 Team members support after end	.344	.374	.347	.224	.522	.529	.530	.424
1.2 Fully explain WA process*	.594	.610	.287	.526	.752	.738	.603	.743
2.1 Written care plan	.488	.500	.295	.338	.628	.636	.462	.445
2.7 Brainstorm strategies*	.667	.710	.352	.529	.793	.820	.559	.744
3.5 Team held responsible*	.545	.564	.405	.490	.720	.726	.597	.681
1.1 Family talk about strengths beliefs	.553	.613	.280	.261	.716	.764	.505	.423
2.11 Understand family beliefs*	.700	.781	.346	.577	.815	.879	.539	.761
1.3 Family talk about what worked	.556	.584	.311	.385	.715	.734	.593	.530
3.12 Involve all members*	.688	.733	.419	.672	.828	.857	.647	.804
3.15 Child communicate ideas*	.423	.417	.295	.468	.655	.618	.551	.619
2.4 Supports connected to strengths*	.589	.647	.489	.386	.787	.810	.740	.647
3.2 Find resources for good ideas*	.500	.580	.303	.269	.645	.719	.451	.474
3.7 Team have new ideas*	.672	.704	.408	.554	.836	.822	.738	.644
3.14 Respect for family	.564	.676	.228	.563	.740	.846	.431	.785
3.11 Positive atmosphere around success*	.623	.680	.422	.583	.810	.845	.687	.758
3.10 Use language family understand*	.307	.327	.144	.337	.524	.534	.336	.522

Note. EFA = exploratory factor analysis; CFA = confirmatory factor analysis; CG = caregiver; Fac.= facilitator; TM = team member; WA = wraparound. Items with asterisks were ultimately removed for shorter 20-item measure.

member scores were so similar as to be virtually identical (Guadagnoli & Velicer, 1991; Lorenzo-Seva & Ten Berge, 2006); three coefficients were above .93, indicating a very high degree of similarity.

Cronbach's alpha coefficients were run for the measure as a whole and for the measure stratified by respondent type. Cronbach's alpha coefficients indicated a good to excellent internal consistency ($\alpha s \geq .90$), although these reliability estimates may be biased due to clustering of sites and respondent types, which is unavoidable in our sample (Waller, 2008).

Although these results do not indicate extremely strong unidimensionality, we considered the data to be "essentially unidimensional" for conducting IRT, which is robust to minor violations in fit (Embretson & Reise, 2000) and still useful even if stronger violations occur (de Ayala, 2009; Harrison, 1986). The pattern of results from the factor analyses (one dominant factor along with several well-correlated minor dimensions) indicate that an IRT using these data would be robust to violations (Kirisci et al., 2001). Additionally, although there are several items with loadings that vary across the sample groups, the analyses of congruence coefficients indicate a very high degree of similarity among sets of factor loadings for each subgroup, providing support for measurement invariance. Therefore, we are confident that an IRT approach is appropriate with this data set and useful for our purposes of item reduction and measurement development, though results should be treated with some caution.

Table 3
Item Locations and Misfit Scores for Initial and Modified Measure

Item	40-item measure				20-item measure			
	Location	Infit[a]	Outfit[a]	DIF[b]	Location	Infit[a]	Outfit[a]	DIF[b]
2.3 Mostly professional services[c]	2.08	1.22	1.45	*ns*				
4.1 Develop transition plan	1.31	1.11	1.20	*ns*	1.21	1.17	1.32	*ns*
3.6 Friend advocate participation	1.25	1.27	1.54	*ns*				
2.5 Strategies get child activities	.95	1.07	1.09	*ns*	.81	1.07	1.07	*ns*
3.3 Child involved activities	.93	1.05	1.08	*ns*	.80	1.04	1.03	*ns*
4.2 Develop youth friendships	.85	.89	.83	*ns*	.71	.88	.84	*ns*
1.4 Family select team members	.52	1.06	1.06	−.54	.34	1.09	1.12	−.50
4.3 Help child solve problems	.45	.89	.83	*ns*	.30	.88	.84	*ns*
3.4 Increase informal support	.44	.88	.82	*ns*	.27	.92	.90	*ns*
4.7 Family succeed without WA	.44	1.25	1.30	*ns*				
3.13 Team end before family ready[c]	.38	1.26	1.40	*ns*				
1.5 Difficult to get team members attend[c]	.33	1.24	1.28	.44				
4.4 Help child prepare for transitions	.28	.89	.80	*ns*	.10	.89	.79	*ns*
2.10 Family not highest priority in design[c]	.17	1.05	1.05	*ns*	−.02	1.12	1.24	*ns*
2.2 Written team vision	.15	1.06	1.15	*ns*	−.04	1.07	1.10	*ns*
3.1 Decisions made family not there[c]	.08	1.01	1.08	*ns*	−.11	1.07	1.23	*ns*
3.9 Team assign review tasks	.05	.90	.75	*ns*	−.14	.90	.81	*ns*
2.8 Crisis or safety plan	.01	1.00	.96	*ns*	−.19	1.01	.98	*ns*
4.6 Family get supportive relationships	.00	.84	.64	*ns*	−.19	.87	.71	*ns*
3.8 Services difficult to access[c]	−.01	1.12	1.14	*ns*				
2.9 Keep child in community	−.01	1.08	.99	*ns*	−.22	1.16	1.12	*ns*
1.6 Identify crises or dangerous situations	−.09	.97	.89	*ns*				
2.6 Members no role implementing[c]	−.15	1.24	1.35	*ns*				
4.5 Team restart if needed	−.23	1.11	1.09	*ns*				
4.8 Team members support after end	−.28	1.07	.92	*ns*	−.50	1.16	1.12	*ns*
1.2 Fully explain WA process	−.40	.89	.67	*ns*				
2.1 Written care plan	−.44	.97	.82	*ns*	−.66	.96	.86	*ns*
2.7 Brainstorm strategies	−.44	.80	.61	*ns*				
3.5 Team held responsible	−.48	.88	.73	*ns*				
1.1 Family talk about strengths beliefs	−.50	.91	.74	*ns*	−.73	.91	.78	*ns*
2.11 Understand family beliefs	−.51	.76	.55	*ns*				
1.3 Family talk about what worked	−.52	.91	.74	*ns*	−.72	.93	.85	*ns*
3.12 Involve all members	−.52	.76	.53	*ns*				
3.15 Child communicate ideas	−.59	1.01	.85	*ns*				
2.4 Supports connected to strengths	−.67	.82	.56	*ns*				
3.2 Find resources for good ideas	−.72	.92	.85	*ns*				
3.7 Team have new ideas	−.78	.77	.43	−.54				
3.14 Respect for family	−.85	.88	.72	*ns*	−1.02	.91	.81	*ns*
3.11 Positive atmosphere around success	−1.03	.81	.45	*ns*				
3.10 Use language family understand	−1.46	.99	.75	*ns*				

Model & measure statistics	40-item measure	20-item measure
Variability accounted for by measure	33.5%	34.2%
Item fit (RMSEA/In MnSq/Out MnSq)	.04 / .99 / .92	.04 / 1.0 / .97
Person fit (RMSEA/In MnSq/Out MnSq)	.36 / 1.0 / .92	.47 / 1.0 / .98
Overall Cronbach's α	.901	.848
Caregiver Cronbach's α	.916	.864
Facilitator Cronbach's α	.830	.755
Team member Cronbach's α	.859	.788

Note. DIF = differential item functioning; WA = wraparound; RMSE = root-mean-square error of approximation; MnSq = mean square.
[a] For Infit and Outfit mean squares, the ideal is 1.0; less than 1.0 indicates overfit, greater than 1.0 indicates underfit. [b] DIF contrasts comparing facilitator with the average of other groups; only the significant *t* tests have displayed contrasts. DIFs for caregivers were nearly exact inverses of those for facilitators. DIFs for other team members were all nonsignificant. [c] Item is reverse scored.

PCMs

After a preliminary Rasch PCM model was run, 36 respondents (1.4% of the sample) were removed from the data set due to grossly misfitting response patterns (*Person Outfit MNSQ* > 2.5). This was generally due to a uniformly high or low response pattern, broken by a few responses in an unpredicted direction on highly discriminating items. Misfitting persons are often removed to improve model and item fit. We chose the seemingly lenient cutoff of 2.5 rather than the more frequently applied 2.0 because our goal was to shorten the measure; hence, we intended to capture as much of the full range of participant responses as possible in order to inform our measure and item estimates, while still removing those people with extreme response patterns indicative of total guessing, complete miscomprehension, or data entry errors. In

other words, we intended to retain persons with misfitting but reasonably possible scores because we wanted their scores to degrade item fit, indicating items that could be problematic. This is in keeping with the Rasch paradigm that data are sacrosanct and important for characterizing item misfit (Andrich, 2004). Nonetheless, there were only an additional 33 respondents (1.4%) with infit or outfit scores above 2.0. This small proportion of extreme scores is expected given a normal distribution of misfit scores, and is too small to have any practical impact on our conclusions.

The model was rerun and results are depicted in Table 3. The model fit Rasch assumptions that randomness is normally distributed as indicated by an approximation to a standardized residual mean of zero and a standard deviation of 1.0. The standardized residual mean was .02 and standard deviation was .97. As a whole, measure scores revealed good model data fit for persons (*Real RMSE* = .36; *Infit MNSQ* = 1.0; *Outfit MNSQ* = .92) and good level of consistency in ordering of person location estimates (*Real reliability* = .80; the word *real* indicates that the estimate has been adjusted for model misfit, hence providing a conservative or "worst case" estimate). In other words, the ordering of persons' probabilities to respond in certain ways based on their estimated level on the latent construct fit the data well, with only a few persons who responded surprisingly different than what the model would predict. More importantly in terms of measure development, results indicated very good model data fit for items (*Real RMSE* = .04; *Infit MNSQ* = .99; *Outfit MNSQ* = .92) and an excellent ordering of item location estimates (*Real reliability* = 1.0).

Table 3 displays the infit and outfit mean squares for individual items. High infit values are indicative of unexpected ratings on items that are located near the person's estimated level of the latent construct (e.g., a pattern of people with average fidelity wraparound teams rating an average-level item with low or high fidelity). High outfit values are indicative of unexpected ratings on items that are estimated to be extremely different than the person's estimated fidelity location (e.g., a pattern of people with high-fidelity wraparound teams rating a high-fidelity item with low fidelity). These values were found to be within an acceptable range. Although various guidelines exist, fit values substantially greater or less than 1.0 are considered indicative of poorly fitting items—one guideline states that scores should fall between .5 and 1.5 (de Ayala, 2009; Linacre, 2009a). Several authors have argued that many recommended fit criteria are much too stringent, emphasizing that practical measurement development is focused on overall measurement quality rather than perfect item fit (de Ayala, 2009; Linacre, 2009a; Pallant & Tennant, 2007). For infit, all items demonstrated mean squares between .76 and 1.27. For outfit, items had mean squares between .43 and 1.54.

In sum, these results indicated that the WFI-4 captures a well-defined unidimensional construct, with items ranked in a consistently predictive way, and with few items that were surprisingly different than the model would suggest. Table 3 displays the ordering of item location. Locations are item-centered, so the average item is located at zero, with high locations indicating high-fidelity items and low locations indicating low-fidelity items. Hence, the Items 2.3 ("Does the family's wraparound plan include mostly professional services"; reverse scored), 4.1 ("Has the team discussed a plan for how the wraparound process will end"), and 3.6 ("Is there a friend or advocate of the child who actively participates on the wraparound team") are items that high-fidelity wraparound teams were more likely to endorse. These items were the least likely to be endorsed by low-fidelity teams. The Items 3.10 ("Do members of the team always use language the family can understand") and 3.11 ("Does the team create a positive atmosphere around success") were answered in a way representing fidelity by the lowest fidelity wraparound teams. Very few respondents did not endorse these items.

DIF. Analysis of respondent-related DIF occurred next. Each group was compared with the average of all the groups and analyzed with a *t* test; the power to detect DIF is very high, so differentially performing items were determined through a combination of the *t* test, the magnitude of difference, and an examination of the graphed item response functions stratified by group (de Ayala, 2009; Linacre, 2009a). The magnitude of DIF contrasts are considered negligible when their absolute values are below .43, slight to moderate when they are between .43 and .64, and moderate to large when they exceed .64 (Zwick, Thayer, & Lewis, 1999).

Table 3 displays the DIF contrast size comparing facilitators with the other groups for those items that were statistically significant. DIF scores comparing caregivers with facilitators and other team members are the nearly exact inverse of these. DIF scores for other team members were all nonsignificant. No items had large DIF. Three items were considered to have slight to moderate DIF: Item 1.4, "Did family members select the people who would be on their wraparound team?" (facilitator contrast size = $-.54$, $t = -9.5$, $p < .001$, implying that at equal levels of wraparound fidelity scores, facilitators were more likely than caregivers and team members to endorse that family members selected persons on the team); Item 1.5, "Is it difficult to get team members to attend team meetings when they are needed?" (facilitator contrast size = .44, $t = 10.5$, $p < .001$, implying that at equal levels of fidelity, facilitators were more likely than caregivers and other team members to report that it was difficult to get team members to attend); and Item 3.7, "Does the team come up with new ideas for the wraparound plan whenever the family needs change or something is not working?" (facilitator contrast size = $-.54$, $t = -4.3$, $p < .001$, implying that at equal levels of fidelity, facilitators were more likely than caregivers and other team members to report that the team comes up with new ideas). Even though multiple tests were run (120 total, 40 for each group), we did not apply a Bonferroni or other correction to our alpha, which would have resulted in unacceptably low levels of power. Because this analysis focuses on identifying items that may be problematic, we were more focused on preventing Type II errors than is typically the case in social science. Nonetheless, the probability of Type I error was still relatively low because the *p* values for significant items were extremely small ($p < .001$), and DIF was considered as only one piece of evidence in a broader analysis of factors pertinent to item retention decisions, as described below.

For several reasons, the last two items (1.5 and 3.7) were deleted from the measure, and Item 1.4 was retained. On the basis of our experience and conversations with practitioners who regularly use the measure, we felt Item 1.4 captured an aspect of central importance to wraparound fidelity (i.e., caregivers and youth actively choose the members of the wraparound team). We believed the moderate DIF for this item reflected an experiential difference

among the respondent types rather than bias (de Ayala, 2009). Additionally, Items 1.5 and 3.7 demonstrated much more discrepant fit scores when compared with Item 1.4, which fit the model very well (see Table 3). After deletion, the model was rerun. The overall person and item fit statistics, item location estimates, and individual item infit and outfit statistics were nearly identical to the 40-item measure.

Category threshold estimates and observed averages. Table 4 shows the observed average statistics and threshold estimates for each category within each item. All of the observed average statistics are in ascending order. Observed average statistics represent the predicted average of the measure estimates that the model produced for people who respond to that category. Because these are in ascending order, it indicates that the categories were arranged in an order that was meaningful to the respondents. For the threshold estimates, there is one less threshold than total number of categories, hence the WFI-4 has three categories and two thresholds. Thresholds should increase in value with category value; disordering indicates one or more of three things have occurred: The category definitions may be out of sequence from how participants understood them (which would be problematic); the category was relatively rarely endorsed; and/or the category defines a narrow section of the response space and could be combined with other categories. Only Items 4.3 and 2.3 had categories that were in ascending order. Hence, the disorder in the threshold estimates is due to either too few respondents endorsing at least one category per item or the category defines too narrow of the response space. Either way, this indicates that nearly all of the items were functioning in a dichotomous fashion and that future measures could change the item response categories into two categories (yes/no) with little loss of information.

Table 4
Observed Average and Threshold Estimates by Category

Item	Observed average (categories)			Threshold estimates (*SE*)	
	0	1	2	1	2
1.1 Family talk about strengths beliefs	.03	.64	1.40	.58 (.09)	−1.59 (.07)
1.2 Fully explain WA process	.06	.58	1.42	.94 (.09)	−1.74 (.07)
1.3 Family talk about what worked	.05	.61	1.40	.83 (.09)	−1.82 (.07)
1.4 Family select team members	.62	.97	1.52	1.72 (.06)	−.71 (.05)
1.5 Difficult to get team members attend	.77	1.05	1.52	.47 (.06)	.16 (.05)
1.6 Identify crises or dangerous situations	.33	.80	1.43	1.42 (.06)	−1.60 (.07)
2.1 Written care plan	.08	.82	1.41	.15 (.09)	−1.05 (.06)
2.2 Written team vision	.49	1.04	1.44	1.21 (.07)	−.93 (.05)
2.3 Mostly professional services	1.08	1.50	1.56	2.01 (.05)	2.04 (.06)
2.4 Supports connected to strengths	−.15	.56	1.42	−.02 (.11)	−1.29 (.07)
2.5 Strategies get child activities	.69	1.31	1.57	1.07 (.05)	.78 (.05)
2.6 Members no role implementing	.70	.97	1.38	1.08 (.08)	−1.41 (.06)
2.7 Brainstorm strategies	−.09	.63	1.44	.39 (.09)	−1.27 (.06)
2.8 Crisis or safety plan	.43	.74	1.45	1.18 (.07)	−1.17 (.06)
2.9 Keep child in community	.37	.95	1.44	.77 (.07)	−.85 (.06)
2.10 Family not highest priority in design	.51	.91	1.47	1.06 (.07)	−.73 (.05)
2.11 Understand family beliefs	−.22	.65	1.44	.17 (.09)	−1.14 (.06)
3.1 Decisions made family not there	.42	.94	1.45	1.04 (.07)	−.90 (.05)
3.2 Find resources for good ideas	−.12	.84	1.42	−.57 (.11)	−.80 (.06)
3.3 Child involved activities	.67	1.32	1.57	1.03 (.05)	.78 (.05)
3.4 Increase informal support	.40	1.05	1.58	.90 (.06)	−.04 (.05)
3.5 Team held responsible	.07	.72	1.43	.23 (.10)	−1.18 (.06)
3.6 Friend advocate participation	.96	1.29	1.54	2.70 (.05)	−.24 (.05)
3.7 Team have new ideas	−.34	.49	1.41	.34 (.11)	−1.79 (.08)
3.8 Services difficult to access	.56	1.01	1.43	.35 (.07)	−.39 (.05)
3.9 Team assign review tasks	.28	.89	1.50	.70 (.07)	−.62 (.05)
3.10 Use language family understand	.06	.59	1.32	−.41 (.16)	−2.22 (.09)
3.11 Positive atmosphere around success	−.50	.47	1.38	−.16 (.13)	−1.78 (.08)
3.12 Involve all members	−.24	.66	1.44	.28 (.10)	−1.30 (.07)
3.13 Team end before family ready	.80	1.13	1.43	1.41 (.06)	−.70 (.05)
3.14 Respect for family	−.16	.58	1.37	.27 (.11)	−1.84 (.08)
3.15 Child communicate ideas	.27	.68	1.39	.30 (.11)	−1.46 (.07)
4.1 Develop transition plan	.88	1.31	1.65	1.79 (.05)	.78 (.05)
4.2 Develop youth friendships	.61	1.17	1.66	1.56 (.05)	.11 (.05)
4.3 Help child solve problems	.40	1.07	1.61	.29 (.06)	.58 (.05)
4.4 Help child prepare for transitions	.38	.93	1.53	1.09 (.07)	−.56 (.05)
4.5 Team restart if needed	.47	.91	1.41	.95 (.08)	−1.42 (.07)
4.6 Family get supportive relationships	.17	.86	1.51	.67 (.07)	−.68 (.05)
4.7 Family succeed without WA	.83	1.11	1.45	.78 (.06)	.04 (.05)
4.8 Team members support after end	.40	.86	1.43	.35 (.09)	−.95 (.06)

Note. WA = wraparound.

Examination of PCM Results to Inform Revision

Though the data fit the models well, both models (i.e., before and after deletion of two items based on DIF results) accounted for only 34% of the raw variability in the data, which is adequate but less than desirable (Linacre, 2009a). This is likely a result of a ceiling effect because most respondents answered very positively. Evidence of this is presented in Figure 1, which depicts the person and item locations, centered by the measure and scaled by logits from −3 to 4, with increasing values representing increasing positive scores on wraparound fidelity. The bar graph on the top provides the number of persons whose score ranks at that level of fidelity. The bar graph on the bottom provides the number of items in the WFI-4 that are located at that level of fidelity.

As shown in Figure 1, the vast majority of the items were located at a level of wraparound fidelity below the response pattern of most respondents. The measure could be improved by the addition of more items targeted on higher fidelity wraparound practice. Additionally, several items assess similar levels of fidelity, as can be seen by the stack of items near zero. This indicates that the measure could be shortened by removing items in this region with little loss of total information. Hence, we explored shortening the measure through an iterative process of removing items in blocks, prioritizing the removal of items with lower and duplicative location estimates and with fit scores below .5 or higher than 1.5. After each iteration, we confirmed the quality of the model.

Analysis of Final 20-Item Measure

The iterative process resulted in 20 items being removed from the measure. Results for the 20-item WFI-4 indicated that very little predictive value was lost. The final model fit Rasch assumptions that randomness is normally distributed, with a standardized residual mean of .01 and a standard deviation of .99. The variability accounted for by the model remained adequate (34.2%). The model data fit for persons was slightly worse for the 20-item measure, but still good (*Real RMSE* = .47, *Infit MNSQ* = 1.0; *Outfit MNSQ* = .98). Scores for person reliability were also slightly less but still good, meaning the ordering of persons' probabilities fit the data fairly well (*Real reliability* = .72). Most importantly, the model data fit for items was excellent and slightly improved (*Real RMSE* = .04, *Infit MNSQ* = 1.0, *Outfit MNSQ* = .97), and retained an excellent, unchanged ordering of item location estimates (*Real reliability* = 1.0).

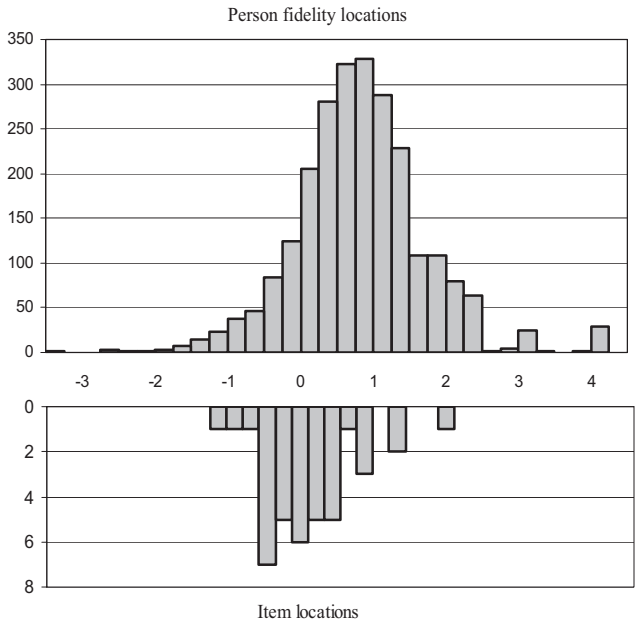

Figure 1. Person- and item-location estimates for the 38-item measure.

Individual item infit and outfit scores are displayed in Table 3. These remained within an acceptable range. DIF estimates remained essentially unchanged—only Item 1.4 had moderate DIF (as in the 40-item model), with a facilitator contrast of −.50, but this item was retained due to reasons described previously.

The estimated latent wraparound fidelity scores, centered by the measure, were examined by each score of the summed 20-item WFI-4. Table 5 presents these data, along with the frequencies and cumulative percentage of people with each score. Latent trait scores increase as total scores increase, but the rate of increase was not linear. This table also depicts the percentile scores for the measure, which illustrate the continued ceiling effect on the measure. Less than 10% of the sample had a score below 20, whereas 50% of the sample had a total score between 33 and 40. At the low end of the scale, less than 1% of the scores have a standard error above .32. At the high end, 15% of the scores have a standard error above .66. This reinforces the need to develop additional items that measure the high end of fidelity in order to increase the reliability of the measure's range of scores. Figure 2 displays the person- and item-location estimates; comparison with Figure 1 indicates a flattening of item-location estimates due to the removal of redundant items.

Internal consistency. Unsurprisingly, given the reduction of items, Cronbach's alpha scores decreased slightly for the measure overall as well as stratified by respondent type, but remained high. Overall alpha for the 20-item measure was .848 (a decrease of .053), caregiver alpha was .864 (a decrease of .052), facilitator alpha was .755 (a decrease of .075), and team member alpha was .788 (a decrease of .071). As indicated earlier, these estimates are likely biased due to the nonrepresentativeness of the sample and the inclusion of different team members in one sample (Waller, 2008).

Discussion

This study was designed to examine and improve the psychometric functioning of the WFI-4 using an IRT modeling approach and to explore the development of a shorter version of the measure to ease participants' response burden while retaining robust measurement of overall wraparound fidelity.

Two items (1.5 and 3.7) were dropped from the measure because they demonstrated different psychometric properties among wraparound facilitators, caregivers, and team members, and because these items demonstrated moderately high misfit with the model. Differences in functioning for Item 1.5, which was related to difficulty in convening wraparound team members for team meetings, are likely due to the fact that wraparound facilitators have the responsibility of gathering team members for meetings, so

Table 5
Correspondence of the WFI-4 20-Item Version to Latent Trait Scores and Frequencies

Score (20-item measure)	Location estimate	Location *SE*	*n*	Cumulative %
0	−3.75	1.74	3	1
1	−2.72	.90	2	1
2	−2.16	.63	3	1
3	−1.85	.51	6	1
4	−1.62	.45	1	1
5	−1.44	.40	4	1
6	−1.29	.37	4	1
7	−1.16	.35	10	1
8	−1.04	.33	10	1
9	−.94	.32	11	1
10	−.84	.31	9	2
11	−.74	.30	12	2
12	−.66	.29	14	3
13	−.57	.29	12	3
14	−.49	.28	14	4
15	−.41	.28	18	5
16	−.33	.28	11	5
17	−.26	.28	22	6
18	−.18	.27	19	7
19	−.11	.27	24	8
20	−.03	.27	28	9
21	.04	.27	33	10
22	.12	.28	43	11
23	.20	.28	33	13
24	.28	.28	47	15
25	.36	.29	56	17
26	.44	.29	71	19
27	.53	.30	75	22
28	.61	.30	77	25
29	.71	.31	105	29
30	.81	.32	101	33
31	.92	.33	134	38
32	1.03	.35	143	44
33	1.16	.37	144	50
34	1.31	.39	158	56
35	1.47	.42	179	63
36	1.67	.47	198	70
37	1.92	.54	157	78
38	2.27	.66	201	85
39	2.87	.94	141	92
40	3.97	1.77	111	97

Note. WFI-4 = Wraparound Fidelity Index, Version 4.

caregivers and other team members may not have the same sense of the effort involved in coordinating the team's schedule. Additionally, it has been proposed that facilitators and caregivers may have different levels of investment in who participates actively in wraparound teams, with facilitators benefiting more from the participation of other professionals who can help coordinate planning and implementation of care plans across child-serving systems (e.g., mental health, schools, and child welfare), whereas parents and caregivers are more interested in convening friends, family members, and community members.

An explanation for differences in Item 3.7—which assesses the team's ability to come up with new strategies for the care plan when things are not working or when the family's needs change—is less clear, but may be due to different understandings of the family's need for change, whether the plan is working, or whether the ideas and strategies being generated are beneficial. Again, this could also be due to different perspectives of professional staff (i.e., facilitators) versus family members, with facilitators possibly viewing formal strategies such as therapy and child welfare services as more important plan components, versus family members and other team members who place greater value on strategies that engage naturalistic sources of support, such as extended family and community resources.

A limitation of this study is its use of a convenience sample collected from participating wraparound sites. IRT models do not require a representative sample in order to calculate valid estimates, only a heterogeneous sample. However, the fact that there is clustering of persons within sites or projects likely resulted in biased model reliability estimates due to uncontrolled-for shared variance. The use of a nonrepresentative sample also places constraints on the development of nationally normed scores. This study is also limited in that there have been no interrater reliability studies done on interviewers' ratings; hence, we have no measurement of the consistency of scores between raters. However, interviewers are trained until they reach at least 80% item-level agreement with "gold-standard" ratings on prerecorded vignettes.

The study is also limited in that the data failed to show strong unidimensionality according to CFA and EFA model fit. However, IRT parameter estimation is relatively unaffected in situations with one dominant factor and several minor dimensions (Kirisci et al., 2001). Additionally, the subgroup factor structures are very similar, according to congruence coefficient testing. Therefore, for our general purposes—item reduction and measurement development—the IRT is useful, but specific results should be interpreted with some caution.

An additional limitation is that the partial credit model of the WFI-4 clearly revealed a ceiling effect, thus resulting in higher measurement error for people with high-fidelity ratings. This likely

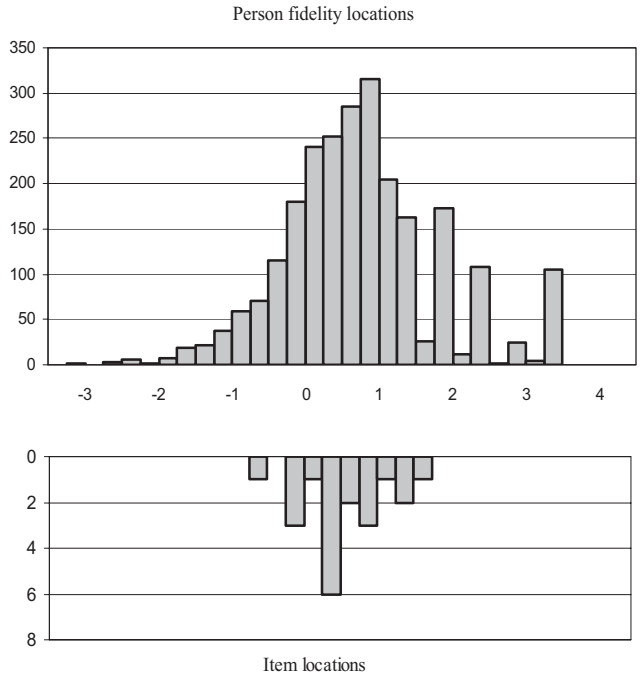

Figure 2. Person- and item-location estimates for the 20-item measure.

reflects several phenomena. First, this may be due to sample characteristics—this was a convenience sample, and participating wraparound sites are likely to be more concerned with fidelity to the wraparound model as compared with nonparticipating providers. Although a representative sample is not necessary when using IRT, future research should focus on teams with a representative range of fidelity in order to accurately specify population norms. In any case, the measure would be improved by the addition of supplementary items capturing the highest levels of wraparound fidelity.

Second, the WFI-4 was developed to evaluate fidelity as conceptualized via the theory- and expert-driven process undertaken by the NWI in 2004. Since that time, wraparound implementation has been increasingly based on this description, with providers, jurisdictions, and states developing training, supervision, and accountability mechanisms to support high-fidelity implementation (Bruns et al., 2011; Walker et al., 2008). Hence, this sample may be more sophisticated in terms of implementing the wraparound process with fidelity than wraparound teams were when this measure was originally developed.

Third, the relative scarcity of high-fidelity WFI-4 items is likely the result of relying on self-report. Some of the WFI-4 items ask respondents for relatively objective or quantifiable information (e.g., the presence or absence of a crisis plan, the relative representation of natural supports on teams, the number of community activities in which the team has successfully engaged the youth). However, many of the items rely on subjective assessments, such as whether the respondent perceives the plan is based on family and community strengths; whether the family is adequately engaged in the process; and whether the wraparound process and the components in the wraparound plan are in line with the beliefs, values, and culture of the youth and family. Although critical to wraparound, these items are more susceptible to demand characteristics and may function similarly to items on satisfaction scales that tend to show ceiling effects (Brannan, Sonnichsen, & Heflinger, 1996; Young, Nicholson, & Davis, 1995).

Results From a 20-Item Model of the WFI-4

Results reveal little change in fit and little loss of information in the measure by the elimination of 20 items, because the eliminated items were redundant in terms of level of fidelity. There was a preponderance of items measuring lower fidelity in both the 40- and 20- item measures. Estimates for most people in the lower range of fidelity are therefore considered quite precise, but the measure lacks precision for the large number of people at the highest levels of fidelity. In regards to other psychometrics, scores for the 20-item WFI were roughly equivalent to the 40-item WFI, but response burden would be cut by half. The internal reliability scores for the shorter WFI decreased only slightly, and were still within the bounds of acceptability. Some decrease in reliability is expected when measures are shortened. These analyses provide some indication that the psychometric properties of the shortened WFI are comparable to the full measure. Further study should confirm these findings using a new sample and extend analyses to test–retest reliability and test validity.

Implications for Measure Development

The findings discussed above raise a basic philosophical question in regards to the construct of fidelity to any practice model. The hypothetically continuous latent construct of "wraparound fidelity" may have an upper limit. This is especially true given the interest in promoting a practice model that is both effective at achieving core outcomes (e.g., maintaining youth at home and in their community and functioning as well as possible despite their challenges) and supported by implementation technologies that are adequate to achieving a level of fidelity necessary for achieving those outcomes. If, in fact, fidelity is associated with positive outcomes (Bruns, Suter, & Leverentz-Brady, 2008; Effland et al., in press), then our primary interest should be to establish fidelity benchmarks regardless of the measure's distributional form. Ultimately, research should focus on determining what implementation technologies most effectively and efficiently promote positive outcomes. Any revisions that are undertaken with the WFI-4 will necessarily need to consider how well it is likely to support such research.

As implied above, use of IRT results to revise a measure such as the WFI-4, which has been found to be useful by many sites nationally, is a challenging venture. Items we suggested to remove may still be considered of importance to practitioners and wraparound teams who may want more detailed information. Fortunately, IRT does not require different test forms to be parallel in order to equate them, meaning that practitioners could use either a long or a short form of the measure and still obtain robust and comparable estimates of wraparound fidelity (Embretson & Reise, 2000), while allowing researchers to include both types of forms in cross-site studies.

Conclusion

Despite these limitations, the current study indicates that the WFI-4 is a robust, unidimensional measure of fidelity to the wraparound process, with an item array across a range of fidelity estimates, and little evidence of gross item bias. However, results of this indicated that the measure could be shortened to 20 items while retaining sufficient psychometric functioning by eliminating items measuring similar levels of fidelity. This could make the measure more useful by reducing time, effort, and expenses in training, administration, response burden, and data entry. Although further studies are needed to replicate current findings and further evaluate the 20-item WFI, we expect these recommendations will be welcomed by practitioners seeking to quickly and efficiently evaluate their wraparound process in order to improve practice, as well as researchers seeking a more efficient method to measure implementation fidelity in randomized comparison trials and other types of research on the wraparound service model.

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implication of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213–232. doi:10.1037/0033-2909.101.2.213

Alexander, J., Barton, C., Gordon, D., Grotpeter, J., Hansson, K., Harrison, R., . . . Sexton, T. (1998). *Blueprints for violence prevention: Book three. Functional family therapy*. Boulder, CO: Center for the Study and

Prevention of Violence, Institute of Behavioral Science, University of Colorado.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1, Supp.), 1–7.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brannan, A. M., Sonnichsen, S. E., & Heflinger, C. A. (1996). Measuring satisfaction with children's mental health services: Validity and reliability of the satisfaction scales. *Evaluation and Program Planning, 19,* 131–141. doi:10.1016/0149–7189(96)00004–3

Bruns, E. J. (2010). *Wraparound Fidelity Index, Version 4: Summary of relevant psychometrics, reliability, and validity studies*. Seattle: University of Washington, Division of Public Behavioral Health and Justice Policy. Retrieved from http://depts.washington.edu/wrapeval/docs/Psychometrics_WFI_April_26_2010.pdf

Bruns, E. J., Burchard, J., Suter, J. C., & Force, M. M. (2005). Measuring fidelity within community treatments for children and families: Challenges and strategies. In M. Epstein, A. Duchnowski, & K. Kutash (Eds.), *Outcomes for children and youth with emotional and behavioral disorders and their families* (pp. 175–197). Austin, TX: Pro-Ed.

Bruns, E. J., Burchard, J. D., Suter, J. C., Leverentz-Brady, K., & Force, M. M. (2004). Assessing fidelity to a community-based treatment for youth: The Wraparound Fidelity Index. *Journal of Emotional and Behavioral Disorders, 12,* 79–89. doi:10.1177/10634266040120020201

Bruns, E. J., Rast, J., Peterson, C., Walker, J., & Bosworth, J. (2006). Spreadsheets, service providers, and the statehouse: Using data and the wraparound process to reform systems for children and families. *American Journal of Community Psychology, 38,* 201–212. doi:10.1007/s10464-006-9074-z

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

Bruns, E. J., Sather, A., & Pullmann, M. D. (2010, March). *The Wraparound Fidelity Assessment System: Psychometric analyses to support refinement of the Wraparound Fidelity Index and Team Observation Measure*. Paper presented at the 23rd Annual Conference of the Research and Training Center on Children's Mental Health, Tampa, Florida.

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,* 726–735. doi:10.1007/s10826-011-9535-3

Bruns, E. J., Suter, J. C., Force, M. M., & Burchard, J. D. (2005). Adherence to wraparound principles and association with outcomes. *Journal of Child & Family Studies, 14,* 521–534. doi:10.1007/s10826-005-7186-y

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

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,* 1586–1593. doi:10.1176/appi.ps.57.11.1586

Bruns, E. J., Suter, J., & Leverentz-Brady, K. M. (2008). Is it wraparound yet? Setting fidelity standards for the wraparound process. *Journal of Behavioral Health Services and Research, 35,* 240–252.

Bruns, E. J., Suter, J. C., Rast, J., Walker, J. S., & Zabel, M. (2006). *Wraparound Fidelity Index, Version 4. Results of an initial pilot test*. Paper presented at the National Wraparound Initiative and Systems of Care meeting, Orlando, Florida.

Bruns, E. J., Walker, J. S., Zabel, M., Matarese, M., Estep, K., Harburger, D., . . . Pires, S. A. (2010). The wraparound process as a model for intervening with youth with complex needs and their families. *American Journal of Community Psychology, 46,* 314–331.

Childs, R. A., & Oppler, S. H. (2000). Implications of test dimensionality for unidimensional IRT scoring: An investigation of a high-stakes testing program. *Educational and Psychological Measurement, 60,* 939–955. doi:10.1177/00131640021971005

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18,* 23–45. doi:10.1016/S0272-7358(97)00043-3

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Domitrovich, C. E., & Greenberg, M. T. (2000). The study of implementation: Current findings from effective programs that prevent mental disorders in school-aged children. *Journal of Educational and Psychological Consultation, 11,* 193–221. doi:10.1207/S1532768XJEPC1102_04

Effland, V. S., Walton, B. A., & McIntyre, J. S. (2011). Connecting the dots: Relationships among stages of implementation, wraparound fidelity, and youth and family outcomes. *Journal of Child and Family Studies, 20,* 736–746

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Epstein, M. H., Nordness, P. D., Kutash, K., Duchnowski, A., Schrepf, S., Benner, G. J., & Nelson, R. (2003). Assessing the wraparound process during family planning meetings. *The Journal of Behavioral Health Services & Research, 30,* 352–362. doi:10.1007/BF02287323

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

Farmer, T. W., & Farmer, E. Z. (2001). Developmental science, systems of care, and prevention of emotional and behavioral problems in youth. *American Journal of Orthopsychiatry, 71,* 171–181.

Forgatch, M. S., Patterson, G. R., & DeGarmo, D. S. (2005). Evaluating fidelity: Predictive validity for a measure of competent adherence to the Oregon model of parent management training. *Behavior Therapy, 36,* 3–13. doi:10.1016/S0005-7894(05)80049-8

Forkmann, T., Boecker, M., Wirtz, M., Eberle, N., Westhofen, M., Schauerte, P., . . . Norra, C. (2009). Development and validation of the Rasch-based depression screening (DESC) using Rasch analysis and structural equation modelling. *Journal of Behavior Therapy and Experimental Psychiatry, 40,* 468–478. doi:10.1016/j.jbtep.2009.06.003

Friedman, R., Katz-Leavy, J., Manderscheid, R., & Sondheimer, D. (1998). Prevalence of serious emotional disturbance: An update. In R. Manderscheid & M. Henderson (Eds.), *Mental health, United States* (pp. 110–112). Rockville, MD: U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration.

Guadagnoli, E., & Velicer, W. (1991). A comparison of pattern matching indices. *Multivariate Behavioral Research, 26,* 323–343. doi:10.1207/s15327906mbr2602_7

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11,* 91–115. doi:10.2307/1164972

Henggeler, S. W., Melton, G. B., Brondino, M. J., Scherer, D. G., & Hanley, J. H. (1997). Multisystemic therapy with violent and chronic juvenile offenders and their families: The role of treatment fidelity in successful dissemination. *Journal of Consulting and Clinical Psychology, 65,* 821–833. doi:10.1037/0022-006X.65.5.821

Hernandez, M., Gomez, A., Lipien, L., Greenbaum, P. E., Armstrong, K. H., & Gonzalez, P. (2001). Use of the system-of-care practice review in the national evaluation: Evaluating the fidelity of practice to system-of-care principles. *Journal of Emotional and Behavioral Disorders, 9,* 43–52. doi:10.1177/106342660100900105

Hoagwood, K., Burns, B., Kiser, L., Ringeisen, H., & Schoenwald, S. (2001). Evidence-based practice in child and adolescent mental health services. *Psychiatric Services, 52,* 1179–1189. doi:10.1176/appi.ps.52.9.1179

Hogue, A., Liddle, H. A., Singer, A., & Leckrone, J. (2005). Intervention fidelity in family-based prevention counseling for adolescent problem behaviors. *Journal of Community Psychology, 33,* 191–211. doi:10.1002/jcop.20031

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit Indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. doi:10.1080/10705519909540118

Kam, C.-M., Greenberg, M. T., & Walls, C. T. (2003). Examining the role of implementation quality in school-based prevention using the PATHS curriculum. *Prevention Science, 4,* 55–63. doi:10.1023/A:1021786811186

Kataoka, S. H., Zhang, L., & Wells, K. B. (2002). Unmet need for mental health care among U.S. children: Variation by ethnicity and insurance status. *American Journal of Psychiatry, 159,* 1548–1555. doi:10.1176/appi.ajp.159.9.1548

Kazak, A. E., Hoagwood, K., Weisz, J. R., Hood, K., Kratochwill, T. R., Vargas, L. A., & Banez, G. A. (2010). A meta-systems approach to evidence-based practice for children and adolescents. *American Psychologist, 65,* 85–97. doi:10.1037/a0017784

Kirisci, L., Hsu, T., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25,* 146–162. doi:10.1177/01466210122031975

Linacre, J. M. (2009a). *A user's guide to Winsteps Ministep Rasch-Model computer programs*. Beaverton, OR: Winsteps.com

Linacre, J. M. (2009b). Winsteps Rasch measurement (Version 3.69.1.7) [Computer software]. Beaverton, OR: Winsteps.com

Lord, F. N., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lorenzo-Seva, U., & Ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 2,* 57–64.

Lyons, J. S., Weiner, D. A., & Lyons, M. B. (2004). Measurement as communication: The Child and Adolescent Needs and Strengths tool. In M. Mariush (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (Vol. 2, pp. 461–476). Mahwah, NJ: Lawrence Erlbaum Associates.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174. doi:10.1007/BF02296272

New Freedom Commission on Mental Health. (2003). *Achieving the promise: Transforming mental health care in America*. Rockville, MD: Department of Health and Human Services.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46,* 1–18. doi:10.1348/014466506X96931

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

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

Sather, A., Cox, K., & Bruns, E. J. (2009). *The Team Observation Measure: Validity, reliability, and psychometrics*. Paper presented at the Building on Family Strengths Conference: Research and Services in Support of Children and Families, Portland, Oregon.

Suter, J. C. (2006). *The wraparound puzzle: Confirmatory factor analysis of the Wraparound Fidelity Index* (Unpublished doctoral dissertation). University of Vermont.

Suter, J. C., & Bruns, E. J. (2009). Effectiveness of the wraparound process for children with emotional and behavioral disorders: A meta-analysis. *Clinical Child and Family Psychology Review, 12,* 336–351. doi:10.1007/s10567-009-0059-y

Tolan, P. H., & Dodge, K. A. (2005). Children's mental health as a primary care and concern: A system for comprehensive support and service. *American Psychologist, 60,* 601–614. doi:10.1037/0003-066X.60.6.601

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. doi:10.1176/appi.ps.57.11.1579

Walker, J. S., Bruns, E. J., & Penn, M. (2008). Individualized services in systems of care: The wraparound process. In B. A. Stroul & G. Blau (Eds.), *The system of care handbook: Transforming mental health services for children and families* (pp. 127–154). Baltimore, MD: Brookes.

Walker, J. S., & Sanders, B. (2011). The Community Supports for Wraparound Inventory: An assessment of the implementation context for wraparound. *Journal of Child and Family Studies, 20,* 747–757.

Waller, N. G. (2008). Commingled samples: A neglected source of bias in reliability analysis. *Applied Psychological Measurement, 32,* 211–223. doi:10.1177/0146621607300860

Weisz, J. R., Jensen-Doss, A., & Hawley, K. M. (2006). Evidence-based youth psychotherapies versus clinical care: A meta-analysis of direct comparisons. *American Psychologist, 61,* 671–689. doi:10.1037/0003-066X.61.7.671

Weisz, J. R., Sandler, I. N., Durlak, J. A., & Anton, B. S. (2006). A proposal to unite two different worlds of children's mental health. *American Psychologist, 61,* 644–645. doi:10.1037/0003-066X.61.6.644

Young, S. C., Nicholson, J., & Davis, M. (1995). An overview of issues in research on consumer satisfaction with child and adolescent mental health services. *Journal of Child and Family Studies, 4,* 219–238. doi:10.1007/BF02234097

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36,* 1–28. doi:10.1111/j.1745-3984.1999.tb00543.x

Appendix A

WFI-4 Items by Principle and Phase

Item	Item text (for wraparound facilitators)	Principle	Phase
1.2	Before first team meeting, did you fully explain the wraparound process and the choices the family could make?	FVC	1
2.10	Would you say that people other than the family have higher priority than the family in designing their wraparound plan?	FVC	2
3.1	Are important decisions ever made about the child/family when they aren't there?	FVC	3
3.15	Does the child have opportunity to communicate his or her own ideas when the time comes to make decisions?	FVC	3
1.4	Did the family select the people who would be on their team?	TB	1
1.5	Is it difficult to get team members to attend meetings when they are needed?	TB	1
2.2	Did the team develop any kind of written statement about what it is working on with the youth and family? AND Can you describe what the team's mission says?	TB	2
3.12	Does the team go out of its way to make sure all members—including friends, family, and natural supports—present ideas and participate in decision making?	TB	3
3.4	Does the team find ways to increase the support the family gets from friends and family members?	NS	3
3.6	Is there a friend or advocate of the child or family who actively participates on the wraparound team?	NS	3
4.2	Has the wraparound process helped the child develop friendships with other youth who will have a positive influence on him or her?	NS	4
4.6	Has the wraparound process helped the family develop or strengthen relationships that will support them when wraparound is finished?	NS	4
2.1	Did the family plan and team create a written plan of care? AND Do they have a copy of the plan?	Col	2
2.6	Are there members of the wraparound team who do not have a role in implementing the plan?	Col	2
2.7	Does the team brainstorm many strategies to address the family's needs before selecting one?	Col	2
3.5	Do the members of the team hold each other responsible for doing their part of wraparound plan?	Col	3
2.5	Does the wraparound plan include strategies for helping the child get involved with activities in his or her community?	CB	2
2.9	Do you feel confident that, in a major crisis, the team can keep the child in the community?	CB	2
3.8	Are the services and supports in the wraparound plan difficult for the family to access?	CB	3
4.7	Do you feel like the child and family will be able to succeed without the formal wraparound process?	CB	4
1.1	Was the family given ample time to talk about strengths, beliefs, and traditions? AND At the first team meeting, were their strengths, beliefs, & traditions shared with all team members?	CC	1
2.11	Did the team take enough time to understand the family's values and beliefs? AND Is the wraparound plan in tune with the family's values and beliefs?	CC	2
3.10	Do members of the team always use language the family can understand?	CC	3
3.14	Do all members of your team demonstrate respect for the family?	CC	3
2.3	Can you summarize the service, supports, and strategies that are in the family's wraparound plan?	Ind	2
2.8	Is there a crisis or safety plan that specifies what everyone must do to respond? AND Does this plan also specify how to prevent crises from occurring?	Ind	2
3.2	When the wraparound team has a good idea for support/services, can it find resources or figure out some way to make it happen?	Ind	3
4.4	Has team helped child prepare for major transitions?	Ind	4
1.3	At the beginning of the wraparound process, was the family given an opportunity to tell you what has worked in the past for the child and family?	SB	1
2.4	Are the supports and services in the wraparound plan connected to strengths and abilities of child and family?	SB	2
3.11	Does the team create a positive atmosphere around successes and accomplishments at each meeting?	SB	3
3.3	Does the wraparound team get the child involved with activities she or he likes and does well?	SB	3
3.13	Do you think the wraparound process could be discontinued before the family is ready for it to end?	Per	3
3.7	Does the team come up with new ideas for wraparound plan whenever the family's needs change AND Does the team come up with new ideas for wraparound plan whenever something is not working?	Per	3
4.5	After formal wraparound has ended, do you think that the process will be able to be "re-started" if the youth or family needs it?	Per	4
4.8	Will some members of the team be there to support the family when formal wraparound is finished?	Per	4
1.6	Before first wraparound team meeting, did you go through a process of identifying what leads to crises or dangerous situations for the child and family?	OB	1
3.9	Does the team assign specific tasks to all members at end of each meeting? AND Does team review each member's follow-through on tasks?	OB	3
4.1	Has the team discussed a plan for wraparound process will end? Does the team have a plan for when this will occur?	OB	4
4.3	Has the wraparound process helped child solve his or her own problems?	OB	34

Note. WFI-4 = Wraparound Fidelity Index, Version 4; FVC = Family Voice and Choice; TB = Team Based; NS = Natural Supports; Col = Collaborative; CB = Community Based; CC = Culturally Competent; Ind = Individualized; SB = Strengths Based; Per = Persistent; OB = Outcomes Based.

(*Appendices continue*)

Appendix B

WFI-4 Item Frequencies

Item	Caregiver			Facilitator			Team member		
	0	1	2	0	1	2	0	1	2
1.1 Family talk about strengths beliefs	116	73	790	27	48	1,132	8	19	181
1.2 Fully explain WA process	142	63	771	18	40	1,127	15	13	160
1.3 Family talk about what worked	121	51	798	23	45	1,118	9	14	168
1.4 Family select team members	431	66	453	58	127	1,025	34	21	140
1.5 Difficult to get team members attend	137	176	662	203	290	729	18	31	167
1.6 Identify crises or dangerous situations	203	52	705	44	56	1,097	18	13	158
2.1 Written care plan	119	118	759	30	85	1,111	9	27	179
2.2 Written team vision	199	116	667	135	60	1,027	13	36	162
2.3 Mostly professional services	591	178	211	636	377	194	110	52	53
2.4 Supports connected to strengths	75	94	799	26	76	1,110	7	17	190
2.5 Strategies get child activities	347	201	369	223	292	585	47	55	103
2.6 Members no role implementing	122	41	790	93	94	1,030	12	14	186
2.7 Brainstorm strategies	136	95	747	19	77	1,121	6	14	195
2.8 Crisis or safety plan	213	94	662	62	66	1,086	19	17	164
2.9 Keep child in community	131	93	730	108	120	951	20	23	168
2.10 Family not highest priority in design	201	93	693	137	127	952	14	31	171
2.11 Understand family beliefs	120	104	768	21	84	1,106	6	22	184
3.1 Decisions made family not there	199	79	706	107	122	1,000	11	24	179
3.2 Find resources for good ideas	77	127	760	15	140	1,060	3	21	191
3.3 Child involved activities	341	202	371	208	291	593	47	57	98
3.4 Increase informal support	274	124	522	114	224	872	33	40	130
3.5 Team held responsible	110	87	752	23	90	1,109	6	21	184
3.6 Friend advocate participation	464	58	470	492	134	594	105	12	98
3.7 Team have new ideas	88	81	789	11	26	1,173	6	12	195
3.8 Services difficult to access	163	145	679	72	181	962	16	32	167
3.9 Team assign review tasks	214	136	606	48	118	1,048	21	30	158
3.10 Use language family understand	32	49	910	6	35	1,118	6	6	207
3.11 Positive atmosphere around success	60	64	860	11	54	1,159	3	11	202
3.12 Involve all members	115	87	775	14	79	1,120	13	16	186
3.13 Team end before family ready	204	89	610	173	105	917	37	32	136
3.14 Respect for family	72	61	860	24	49	1,154	7	10	203
3.15 Child communicate ideas	89	76	716	18	50	1,034	5	19	187
4.1 Develop transition plan	520	169	281	346	255	614	93	36	77
4.2 Develop youth friendships	340	103	442	207	196	683	68	33	102
4.3 Help child solve problems	230	248	422	90	266	738	23	76	112
4.4 Help child prepare for transitions	255	105	548	76	110	886	21	37	134
4.5 Team restart if needed	99	49	724	79	71	1,032	15	21	156
4.6 Family get supportive relationships	154	94	715	84	146	984	30	34	145
4.7 Family succeed without WA	187	131	631	174	235	797	43	57	101
4.8 Team members support after end	104	76	736	55	114	1,029	15	36	158

Note. WFI-4 = Wraparound Fidelity Index, Version 4; WA = wraparound.

Received September 29, 2010
Revision received January 2, 2013
Accepted January 14, 2013 ■