HHS Public Access

Author manuscript

Psychol Assess. Author manuscript; available in PMC 2016 June 01.

Published in final edited form as:

Psychol Assess. 2015 June ; 27(2): 710–725. doi:10.1037/a0038555.

Therapist Perception of Treatment Outcome: Evaluating Treatment Outcomes among Youth with Antisocial Behavior Problems

Brent R. Crandal,
Chadwick Center for Children and Families, Rady Children's Hospital – San Diego

Sharon L. Foster,
Alliant International University - San Diego

Jason E. Chapman,
Medical University of Southern Carolina

Oregon Social Learning Center.

Phillippe B. Cunningham,
Medical University of Southern Carolina

Patricia A. Brennan, and
Emory University

Elizabeth A. Whitmore
University of Colorado Denver School of Medicine.

Abstract

Effective evaluation of treatment requires the use of measurement tools producing reliable scores that can be used to make valid decisions about the outcomes of interest. Therapist-rated treatment outcome scores that are obtained within the context of empirically supported treatments (EST) could provide clinicians and researchers with data that are easily accessible and complimentary to existing instrumentation. We examined the psychometric properties of scores from the Therapist Perception of Treatment Outcome: Youth Antisocial Behavior (TPTO:YAB), an instrument developed to assess therapist judgments of treatment success among families participating in an EST, Multsystemic Therapy (MST), for youth with antisocial behavior problems. Data were drawn from a longitudinal study of MST. The initial 20-item TPTO was completed by therapists of 111 families at mid-treatment and 163 families at treatment termination. Rasch model dimensionality analyses provided evidence for two dimensions reflecting youth- and caregiver-related aspects of treatment outcome, although a bifactor analyses suggested that these dimensions reflected a single more general construct. Rasch analyses were also used to assess item and rating scale characteristics and refine the number of items. These analyses suggested items performed similarly across time and that scores reflect treatment outcome in similar ways at mid and post-treatment. Multilevel and zero-order analyses provided evidence for the validity of TPTO scores.

Correspondence concerning this article may be addressed to Brent Crandal, PhD, Chadwick Center for Children and Families, Rady Children's Hospital, 3020 Children's Way, MC 5131, San Diego, CA 92123. bcrandal@rchsd.org.

TPTO scores were moderately correlated with scores of youth and caregiver behaviors targeted in treatment, adding support to its use as a treatment outcome measurement instrument.

Keywords

assessment; treatment outcome; empirically supported treatment; Multisystemic Therapy; antisocial; youth

Evaluation of the efficacy of psychological intervention depends on the ability to measure treatment outcomes in psychometrically sound ways, using instruments that draw from multiple perspectives. Unfortunately, the family-based treatment literature for externalizing disorders has suffered from a lack of consistency in the measurement of treatment success (i.e., outcome). In a review of 1,430 instruments used in professional journals over a 5-year period, Froyd, Lambert, and Froyd (1996) found that approximately 60% of therapy outcome measurement tools were used only one time, and often their psychometric properties were not appropriately evaluated. Researchers have offered several sets of guidelines to improve measurement of treatment outcome, ranging from multicultural assessment considerations (e.g., Lau et al., 2004; Murphy, Faulkner, & Behrens, 2004) to recommendations of types of treatment outcomes that should be assessed (e.g., Kazdin, 1979; Rosen & Procter, 1981). Despite some attention to types of data sources or raters (Kazak et al., 2010; Leibert, 2006) and the accepted importance of multi-source, multi-method assessment, these guidelines have largely neglected the perspective of treatment providers in assessing outcomes. In this study we describe the development and initial evaluation of the Therapist Perception of Treatment Outcome: Youth Antisocial Behavior (TPTO:YAB), a new measurement tool developed to assess therapist perceptions of treatment success for families participating in a widely-used EST, Multsystemic Therapy (MST), for youth with antisocial behavior problems. Youth antisocial behaviors include a wide variety of externalizing conduct problems, including aggressive, disruptive, delinquent and related behaviors, and are among the most common problems reported by those seeking child and adolescent treatment (e.g., Ebesutani, Bernstein, Chorpita, & Weisz, 2012).

Outcome Assessment Considerations

Several factors make empirically supported treatments (ESTs) for youth antisocial behavior problems a useful context in which to develop new psychometrically-sound treatment outcome assessment instruments completed by treatment providers. First, the highly specified treatment goals and methods of ESTs for youth antisocial behavior problems offer frameworks for the content of treatment outcome measurement tools. Second, new treatment outcome assessment tools that are developed in the context of specific ESTs (such as MST) can often be linked to other well-established outcome and treatment process data that are routinely collected in these contexts. Third, the evaluation of treatment outcome from multiple perspectives has been an important research topic in the evidence-based practice movement with antisocial youth (Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2002). Within this context, new treatment outcome measurement tools that are consistent with treatment goals, informed by a unifying theoretical framework, and that can

be readily incorporated into ongoing assessment protocols have the potential to be integrated into both research and practice contexts.

In their exhaustive review of psychological assessment, Meyer et al. (2001) concluded that multiple respondents provide the best way to assess relevant clinical factors throughout the course of treatment. Although caregiver, youth, and teacher perceptions are routinely assessed, therapist perceptions are less frequently measured and few assessment tools have been developed to gather therapist perspectives on treatment outcome. For example, a comprehensive review of assessment measures for conduct problems did not mention a single therapist questionnaire instrument (McMahon & Frick, 2007). Although some measurement tools have been adapted for use by therapists (i.e., the Youth Outcome Questionnaire: Omni Version; Dunn et al., 2005, and the Child Behavior Checklist; Dutra, Campbell, & Westen, 2004), the evidence for instrument performance with therapist respondents on these measures has not been sufficiently evaluated to draw meaningful conclusions.[1] In other cases, therapist-rated measures have been developed solely to standardize the therapist decision-making process related to transfers from juvenile courts or determining risk (e.g., the Risk-Sophistication-Treatment Inventory; Salekin, 2005, and the Youth Level of Service/Case Management Inventory; Hoge & Andrews, 2002; Hoge, Andrews, & Leschied, 2002). Therapist judgments about treatment outcome (as opposed to symptom counts) are not distinctively targeted in either case.

Therapists providing ESTs are generally trained to set goals and follow specific treatment procedures, and are important treatment partners in determining when families should terminate. Therapists also engage in ongoing clinical evaluation of how the family is responding to treatment throughout therapy. A psychometrically-sound evaluation of treatment outcome that captures the ways therapists judge family progress in ESTs could capitalize therapists' experience and training, as well as their access to ongoing, real time information.

The purpose of the current study was to evaluate evidence for the reliable and valid use of scores from a new therapist measurement tool, the TPTO:YAB, designed to capture therapist judgments regarding treatment outcome among families participating in MST, an evidence-based family therapy designed to reduce antisocial behaviors in youth. This therapist-report instrument was designed to be used to assess youth, caregiver, and overall family treatment success in the context of family-based EST.

Multisystemic Therapy

Multisystemic Therapy is a widely disseminated treatment for antisocial behaviors (e.g., delinquency, substance abuse) in youth that has generated substantial evidence of its effectiveness (Kazdin & Wassell, 1998). Since the first quasi-experimental trial of MST (Henggeler et al., 1986), more than 20 independent reviewers have supported MST as an

[1]Dutra et al. (2004) assessed the psychometric properties of the CBCL as completed by 294 clinicians who reported on behaviors of adolescent clients. Although Dutra et al. concluded that clinicians can provide data with acceptable levels of reliability and validity, the conclusions were based on comparison of therapist-reported data with therapist report on the parent version of the CBCL, without the benefit of additional respondent perspectives on treatment outcome.

evidence-based treatment for serious antisocial behavior (Henggeler et al., 2009) and some consider it to be a model program (e.g., Tate & Redding, 2005). Desired outcomes of MST, as with other ESTs for antisocial youth, include (a) reduced criminal and drug use behaviors, decreased psychiatric symptomology and externalizing problems, (b) improved family interactions and parenting, and (c) improved youth functioning with peers and in school. These outcomes are clearly emphasized in the MST treatment manual (Henggeler et al., 1998) and are key targets for MST interventions. Thus, MST provided an excellent context in which to develop and evaluate a tool that assessed therapist perspective of treatment outcome.

Goals of the Study

The TPTO:YAB was designed to supplement caregiver and youth report measures typically used to assess the effects of ESTs for antisocial behavior problems by providing a therapist-completed measure that provides a quantitative index of the kinds of judgments therapists make in determining whether their clients are successfully completing treatment. One primary goal of this study was to determine whether TPTO:YAB scores provided reliable indices of treatment outcome for diverse families of antisocial youth receiving MST in community mental health settings. Towards this end, we first examined TPTO:YAB item and rating scale functioning. We also examined the internal structure of the TPTO:YAB, and explored whether results were consistent across two administrations—one at mid-treatment and one at the end of treatment.

Additional goals involved examining validity evidence for the TPTO:YAB scores based on correlations between TPTO:YAB scores and scores from established measurement instruments assessing other domains of youth and family functioning consistent with the MST theory of change (Henggeler et al., 1998). Specifically, we looked at whether TPTO:YAB scores were associated with scores assessing caregiver and youth reports of youth externalizing behavior outcomes. Because improved parenting plays an instrumental role in MST (Henggeler et al., 2009; Henggeler & Schaeffer, 2010; Huey, Henggeler, Brondino, & Pickrel, 2000), we also examined correlations between TPTO:YAB scores and scores based on caregiver ratings of (a) discipline and monitoring, and (b) feelings of incompetence and guilt. Furthermore, we expected TPTO:YAB scores to predict group classification of successful versus unsuccessful treatment responders, based the circumstances for therapy discharge. Because youth externalizing behaviors are a primary focus of MST and internalizing behaviors are not necessarily targeted, we also expected associations between TPTO:YAB scores and caregiver reports of youth internalizing behaviors to be weaker than correlations between TPTO:YAB scores and scores assessing youth externalizing behaviors. Finally, we examined whether TPTO:YAB scores reported at mid-treatment significantly predicted outcomes at termination.

Methods

Design and Procedures

Families involved in a longitudinal evaluation of MST in real-world practice settings provided data for the study. Procedures were approved by authors' Institutional Review

Boards for the Protection of Human Participants, and therapists, youth, and caregivers signed consent forms upon enrollment in the study. Youth and their caregivers completed computerized assessments at five time points: early in treatment (Time 1; T1), twice during mid-treatment (T2 and T3, after about 6-8 and 12-14 weeks of treatment, respectively), at treatment termination (T4), and at six-month follow-up (T5). Youth and caregiver data from T1, T3, and T4 assessments were used in analyses reported here. Therapists completed the TPTO:YAB as part of a computerized assessment battery at two time points: during mid-treatment (T3) and at treatment termination (T4).

Participants

Forty-four therapists provided TPTO:YAB data in the current study. The majority (72.7%) was female; 84.1% self-identified as White, 6.8% as Spanish, Hispanic, or Latino, 4.5% as Asian, 2.3% as Black or African-American, 2.3% as more than one race, and 2.3% as "Other." Therapists had spent on average 9.6 months providing MST at the time they enrolled in the study. Most (79%, $n = 35$) had a Master's degree, 13% ($n = 6$) had a college degree, and 6.8% ($n = 3$) had a doctoral degree. All therapists had completed standard MST training (Henggeler et al., 2009) as a requirement for participation. The median number of families rated by any therapist was 3 (range, 1-13).

Families ($N = 185$) receiving MST were recruited for the parent study from four licensed MST programs in the Denver, Colorado area. Youth and families were referred primarily from social service agencies and the juvenile justice system. Inclusion criteria for the larger study were: (a) families with a son or daughter between 12 and 17 years old referred for MST services based on serious externalizing behavior problems, (b) had been living in the caregiver's home for a month or more at the time of referral with no immediate plans for out-of-home placement, and (c) had one or more caregivers willing to participate in the study. On average, families who participated in the study spent 17.5 weeks (5 months) in treatment with a range of 3 to 43 weeks.

The sample for the current study consisted of the 163 families[2] whose therapists provided youth or caregiver reports on the TPTO:YAB at T4. These families included 104 (63.8%) male and 59 (36.2%) female youth participants. Over 47% ($n = 78$) of the participating youth identified as White, 29.4% ($n = 48$) identified as Spanish, Hispanic, or Latino, 18.4% ($n = 30$) identified as Black or African-American, and less than 1% ($n = 1$) identified as Asian or American Indian/Alaska Native ($n = 1$). Four participants (2.5%) identified as more than one race and one participant (.6%) responded as Unknown. Youth averaged 15.4 ($SD = 1.3$) years of age (range 12-17). Almost all (98.2%) of the caregiver respondents considered themselves to be the primary caregiver (76.7% were the youth's biological mother, 9.2% youth's father, 8% youth's grandmother, 6.1% self-identified as "Other"). Of these caregivers, 47.9% had not earned a high school degree, 12.3% highest level of education was a high school degree, 38.0% had completed some college or graduated, and 1.8% had a graduate level education. In addition, 41.6% of caregivers reported that the family was receiving financial assistance.

[2]The TPTO:YAB was not IRB approved as an addition to the study until after the first 22 families had completed the study.

TPTO:YAB data were available at mid-treatment (T3) for 111 families. Missing data at T3 were principally due to families terminating prior to reaching the mid-treatment assessment (families who terminated early received T4 assessment at termination in place of the scheduled T3 assessment). Data at T3 and T4 were also missing due to occasional difficulty locating and scheduling a family, a family member declining an assessment, and technical problems with computers used to administer measures. Families who had missing data at T3 and T4 were evaluated to see if their symptom severity at T1 and demographic characteristics distinguished them from those included in this study. No statistically significant differences between the groups emerged on caregiver report of pre-treatment youth externalizing behaviors on the Child Behavior Checklist, or youth and caregiver age, ethnicity or gender.

Measurement Tools

The Therapist Perception of Treatment Outcome: Youth Antisocial Behaviors —(TPTO:YAB) was designed to assess therapists' perception of treatment success among families they treat. The TPTO:YAB was intended to complement and not duplicate existing measures of outcome commonly used with ESTs for antisocial behavior, and to capture the ways therapists judge positive response to treatment. Although principally intended as an assessment of outcome at termination, therapists completed the TPTO:YAB at T3 and T4, permitting a quasi-replication of T4 results using T3 data. We evaluated psychometric properties of TPTO:YAB scores at both time points.

TPTO:YAB items were developed based on semi-structured qualitative interviews of nine therapists recruited from different MST sites from across the U.S. These therapists averaged 34 years of age (range, 25 - 60), with an average of 2.2 years of experience providing MST (range, 8 months - 6.5 years). Seven were female; six were White, two African-American, and one Puerto Rican. Three were Bachelor's level clinicians and the remaining six had Master's degrees. During 30-minute interviews, a graduate student interviewer asked open-ended questions designed to elicit characteristics of the family and MST treatment process that the therapist believed distinguished families who responded well to MST from those who did not.

Three of the authors reviewed the therapists' verbatim responses and generated items in accord with item content selection procedures (Clark & Watson, 1995; Haynes et al., 1995; Vogt, King, & King, 1995). Specifically, they first independently created lists of salient themes that emerged from the qualitative interviews with the MST therapists. Next, they collaboratively identified 16 distinct general themes (e.g. improved caregiver communication across systems, improved youth functioning, and caregiver demonstrations of engagement, problem solving, and generalization of treatment skills) based on review of the collective list and consolidation of overlapping themes. The authors then each generated 32 items based on aspects of the 16 themes, identified representative items from the collective item pool, and obtained feedback on wording from the remaining authors.[3] The

[3]Due to damaged electronic files, the final 16 themes and initial 32 items created while developing the TPTO:YAB are no longer accessible to the authors. These data could assist with further qualitative evaluation of the development process and the absence of these data is a limitation to further in-depth examination of the validity of item content.

final 20 items were selected for the total scale (see Table 1) based on consensus that they tapped unique aspects of the themes, were not redundant with items in other instruments used in the study, and were worded clearly. Each item was rated using a 6-point rating scale (i.e., agree strongly; agree; agree slightly; disagree slightly; disagree; disagree strongly).

Alabama Parenting Questionnaire (APQ)—The APQ (Frick, 1991) assesses parenting practices associated with disruptive behavior in youth. The caregiver-completed APQ poor monitoring and inconsistent discipline subscale scores were used in this study, as these constructs are theoretically consistent with MST treatment goals and these scores would be expected to correlate with TPTO:YAB scores. Hawes and Dadds (2006) found good evidence for construct validity of the APQ subscale scores among youth diagnosed with Oppositional Defiant or Conduct Disorder. In the current study coefficient alphas for the inconsistent discipline and poor monitoring/supervision items were .76 and .79 at T3, and .72 and .79 at T4, respectively.

Stress Index for Parents of Adolescents (SIPA)—The SIPA (Sheras, Abidin, & Konold, 1998) is a 112-item self-report tool that assesses parenting stress for caregivers of adolescents. Construct validity of SIPA subscale scores has been supported by significant positive correlations with CBCL Externalizing and Personality Assessment Screener scores (Morey, 1997). In this study, we used raw scores on the incompetence/guilt (INC) subscale, as this subscale content most closely aligns with MST treatment goals. The INC scale contains items measuring how confident the parent is about coping with the youth and the presence of guilt feelings in different situations (e.g. when the adolescent misbehaves or gets in trouble). Internal consistency of the INC scale items in the current study was .86 at T3 and .85 at T4.

Child Behavior Checklist—The CBCL (Achenbach, 1991) is one of the most frequently used measurement tools examining child behavioral functioning. The CBCL consists of 113 behavior problem items and includes three broadband behavior problem scales (Internalizing, Externalizing, and Total Behavior Problems). Sawyer et al. (1990) found that all quantitative scale scores significantly discriminated between referred and non-referred children. In this study, the externalizing and internalizing scale raw scores were used, with higher scores indicating more problems. Although externalizing and internalizing behaviors are often significantly associated, we expected that TPTO:YAB scores would be more strongly related to externalizing behavior (which are a primary target in MST) scores than internalizing behavior scores. Coefficient alphas for the raw score subscale items at T3 were .95 (Externalizing) and .90 (Internalizing). At T4, the coefficient alphas were .95 (Externalizing) and .91 (Internalizing).

Case Discharge Summary—We used the Case Discharge Summary (CDS; Schoenwald, Sheidow, Letourneau, & Liao, 2003) to group participants by successful vs. unsuccessful termination at outcome. The CDS format requires therapists to select a reason for discharge (e.g., MST goals completely met; some goals met but diminishing returns for treatment; youth placed out of home during treatment; family requested treatment termination; referral source closed case; reimbursement source closed case; therapist/team requested termination)

and to identify (from a list of 15 options) who closed the case (Schoenwald et al., 2003). These ratings were coded dichotomously. Families who were rated as having met treatment goals, and who had terminated based on agreement between the therapist and the family, were grouped as the treatment successes. Remaining families were considered the treatment non-success group. In Schoenwald et al.'s MST transportability study, CDS scores were significantly predicted by scores of therapist adherence during treatment and correlated with scores on other instruments in the anticipated directions, supporting the validity of CDS scores.

Self-Report Delinquency Scale (SRD)—The SRD (Elliott et al., 1983; Elliott et al., 1985) is among the best-supported of the self-report delinquency scales (Henggeler, 1989). The 47 items of the SRD assess covert and overt antisocial behavior. Items inquire about how many times youth have committed a specific offense within the last 90 days. Huizinga and Elliott (1986) provided good evidence for construct validity of scores based on comparisons of SRD scores and arrest records. The general delinquency score was used in the current study, with scoring based on Rasch analyses (Rasch, 1960, 1980; Chapman, personal communication, 2010) that indicated that dichotomized items provided better indicators than frequency scores, primarily due to the skewed distribution of SRD item scores. Item scores were dichotomized and the number of items endorsed was summed. Internal consistency of SRD items in the current study was found to be .41 and .77 at T3 and T4 respectively. Although the internal consistency for SRD items at T3 was lower than expected, we used T3 data because of the importance of assessing youth report of externalizing behaviors, and interpreted T3 results with caution.

Analytic Approach

Rasch Modeling—Rasch modeling (Rasch, 1960, 1980; Bond & Fox, 2007) is an increasingly used contemporary approach to item analyses and offers ways of assessing the rating scale and item characteristics of the TPTO:YAB not addressed by conventional psychometric approaches (e.g., internal consistency analyses). The Rasch model has traditionally been used to score and evaluate tests comprised of items with correct/incorrect (i.e., dichotomous) response formats (e.g., achievement tests). According to the dichotomous Rasch model, when a person responds to an item, the probability of a correct response is the net result of the person's ability and the item's difficulty. For example, given a person with high ability and an item with low difficulty, the probability of a correct response is high. The polytomous Rasch model is used to accommodate items with rating scale (i.e., ordered categorical) responses, providing a highly flexible model for evaluating the performance of measurement instruments such as the TPTO:YAB. The Rasch rating scale model provides an evaluation of dimensionality in the data and rating scale performance, interval scale measures (i.e., scores) and fit statistics for each item and person, and reliability statistics for the sample of items and persons. For the present study, the models were performed using WINSTEPS software (Linacre, 2013). Prior to analysis, items were reverse coded as needed so that a higher response reflected the perception of a positive outcome. Separate models were performed for the T3 and T4 data.

Multilevel Analysis—Many therapists provided data on more than one family in this study and thus data could not be considered independent. Intraclass correlations (ICCs) for TPTO:YAB items calculated using MPlus v. 6.11 (Muthén & Muthén, 1998-2010) indicated that these relationships were nontrivial: at T3 these ranged from .27-.53, and at T4 from .14 - .37. Similarly, intraclass correlations (ICCs) for subscale and total scores (see Results) were also high: total score ICC = .34 and .29, ECO scale = .22 and .21, and YBO scale = .28 and .26 at T3 and T4, respectively. ICCs higher than .10 are commonly considered suggestive of problems with independence (Cohen et al., 2003). Therefore, multilevel analyses (in which scores for families were nested within therapists) were used to examine data whenever possible.

Traditional Validation Approaches—The results of the Rasch measurement model provide a number of sources of evidence for evaluating the reliable and valid use of an instrument's scores; however, additional validity evidence is provided by using those scores to predict the extent to which scores generated by an instrument relate in meaningful ways to scores from tools assessing related constructs or to whether scores on new instruments can be used for their intended purposes. To address these issues, we drew from traditional validation approaches. Specifically, if the TPTO:YAB works as intended, then the scores should correlate with different treatment outcome validation tool scores and TPTO:YAB scores were used to examine evidence for construct-related validity using two-level Mplus analyses to estimate family-level (within) correlations separated from therapist-level (between) effects. Multilevel logistic regressions were used to assess whether scores significantly differed for two participant groups classified as treatment success and treatment non-success at termination. Finally, multilevel linear and negative binomial regression analyses provided information regarding whether TPTO:YAB scores collected in mid-treatment predicted outcome at termination.

Results

Rating Scale Performance and Dimensionality

The first goal of the study was to examine TPTO:YAB item and rating scale functioning to inform decisions about how to score the TPTO:YAB. To determine whether therapists used the 6 points of the rating scale consistently and as expected, we evaluated the category structure of the TPTO:YAB 6-point rating scale using methods described by Linacre (2002). The results using all 20 items at T4 indicated that the rating scale did not function as intended. Although the rating scales were used monotonically, evidence suggested that therapists did not discriminate well between the two middle categories (i.e., Agree Slightly, Disagree Slightly). Step calibrations (which are expressed as logits) reflect the distance between adjacent categories, with a recommendation of at least 1.00 logit between categories for a rating scale with five categories. Categories 3 and 4 were separated by only 0.30 logits, and categories 4 and 5 were separated by 0.67 logits. Results at T3 were comparable. Therefore, to optimize the scales, the two middle categories were combined, reducing the 6-point scale to a 5-point scale.

The Rasch models used to assess rating scale functioning assume that items used in analyses assess a unidimensional construct. Rasch dimensionality assessments address this assumption. A dimensionality assessment based on all 20 items with responses on the 5-point rating scale indicated that of the total variance in the T4 TPTO:YAB reports, 64.2% was explained by the Rasch item and person measures. Using the remaining pool of unexplained variance, dimensionality was evaluated by performing a principal components analysis (PCA) on the standardized Rasch item residuals after the first dimension had been extracted (Smith, 2002). The first contrast (i.e., the component explaining the greatest proportion of the residual variance) had an eigenvalue of 5.4, suggesting that there was likely meaningful dimensionality in the data (i.e., eigenvalue ≥ 2.0; Linacre, 2013). Inspection of the item loadings on this contrast revealed a clear pattern of results. Specifically, one dimension was formed by the items that referenced caregivers (i.e., 2, 4, 7, 9, 11, 12, 13, 14, 15, 16, 17, 19, 20) and the second by the items that referenced the youth or family (i.e., 1, 3, 5, 6, 8, 10, 18). Accordingly, the items were divided along these lines. Separate models were performed to evaluate whether non-trivial dimensionality was present within these dimensions. For the Caregiver dimension, 68.8% percent of the variance was explained by the model and the eigenvalue for the first contrast was 2.4; and for the youth dimension, 71.1% of the variance was explained, with an eigenvalue of 1.7 for the first contrast. These results suggest no substantial dimensionality within the Caregiver or Youth dimensions. For T3 data, the conclusions were highly consistent. Only two items (15, 19) loaded on the opposite dimension; however, across both the T3 and T4 data, these items loaded weakly on the respective dimensions.[4]

As a check on the findings regarding rating scale functioning, we repeated the analyses with the 6-point scale for the Caregiver and Youth dimensions separately. For the Caregiver dimension, categories 3 and 4 were separated by only 0.12 logits, and categories 4 and 5 were separated by only 0.08 logits. The finding was similar for the Youth dimension. In contrast, the optimized 5-point rating scale performed well for both the Caregiver and Youth dimensions, with at least 1.81 and 1.84 logits, respectively, between adjacent categories at T4. This coding strategy was used for all of the results reported subsequently. For the T3 data, the results were the same.

[4]Rasch dimensionality assessments did not take into account the multilevel nature of the data. Therefore, we conducted a set of multilevel exploratory factor analyses (EFAs) of 6-point TPTO:YAB items at T4 and T3 to supplement the Rasch dimensionality analyses. EFAs used an oblique (geomin) rotation and used Mplus Version 6.11 (Muthén & Muthén, 1998-2010), with families nested within therapist. Up to four factors were tested at the within-subjects (family) level. An unrestricted model was specified at the therapist level because: (a) the focus of the study was on family-level results, not on the factor structure for therapists collapsed across families, and (b) the number of therapists was too small for valid estimates at the therapist level. To determine the appropriate number of factors, we examined model fit using the Akaike Information Criterion [tidelAIC] index calculated as part of the multilevel analyses, supplemented with parallel analysis (PA; calculated using SPSS). Ruscio and Roche (2012) reported that AIC and PA approaches had 73% and 76% accuracy in identifying the appropriate number of factors in their Monte Carlo study, respectively. At T4, AICs decreased as additional factors were added (AICs = 9060.61, 8460.41, 8409.59, 8375.37 for 1, 2, 3, and 4-factor solutions, respectively). However, parallel analysis indicated that a two-factor structure was most appropriate. At T3, only 1 and 2 factor solutions could be reliably calculated with the multilevel EFA; parallel analysis again supported a 2-factor solution. Consistent with dimensionality results, one factor contained items generally focused on youth treatment outcomes (Items 1, 3, 5, 8, 10, and 18). The other set of items addressed caregiver treatment outcomes (Items 2, 4, 7, 9, 11, 12, 13, 14, 16, 17, and 20). Three items cross-loaded on both scales (Items 6, 15, and 19).3 T3 and T4 results were consistent. Tables of factor loadings for multilevel EFAs with the T3 and T4 items with 6-point and 5-point scoring are available from the first or second author upon request.

Item Selection

Items were further evaluated with regard to their subscales based on item fit statistics.[5] Item fit was evaluated using standardized outlier-sensitive fit statistics (i.e., outfit) with a critical value of ≥ 2.0 (Smith, 2000; see Tables 2 and 3). This statistic identifies items characterized by unpredictable responses, such as an item not being endorsed strongly by therapists who would be expected to endorse it strongly. Across dimensions, several items were identified as significantly misfitting. For the Caregiver dimension, this included items 9 (outfit = 3.3), 11 (outfit = 5.5), 16 (outfit = 2.8), and 19 (outfit = 2.3), and for the Youth dimension, items 6 (outfit = 2.8) and 18 (outfit = 6.4). The misfitting items were reviewed for content and removed in a stepwise fashion based on the level of misfit. The final model for the Caregiver dimension included nine items (2, 4, 7, 9, 12, 13, 14, 17, 20) and for the Youth dimension included five items (1, 3, 5, 8, 10). For T3 data, the results were generally consistent: there were three significantly misfitting items in the Caregiver dimension, including 11 (5.7), 15 (2.1), and 16 (2.1). For the youth dimension, one item, 18 (2.3), showed misfit. The final items were selected based on the T4 data, and in the resulting models no items were misfitting on the Caregiver or Youth dimensions.

Item and Person Reliability

Across dimensions, two reliability statistics, as detailed by Smith (2001), were evaluated for the sample of items and the sample of persons. The first is similar to traditional reliability estimates, with values that can range from 0 to 1. In all cases, the estimates at T4 were high. For the Caregiver dimension, the reliability for the sample of families (person reliability) was .94 and for the sample of items was .95. For the Youth dimension, person reliability was .94 and the item reliability was .97. These values suggest that reliability is quite high, with only a small proportion of variance attributable to measurement error. Person (Caregiver dimension = .93; Youth dimension = .87) and item (Caregiver dimension = .94; Youth dimension = .96) reliability values at mid-treatment were also high.

The second reliability statistic, separation reliability, can range from 0 to infinity. For person separation, the value reflects the number of meaningfully distinctions that can be made in the sample of persons (i.e., in the level of perceived treatment outcomes) by using the sample of items. The reverse is true for item separation, with the value reflecting the number of distinctions in items that can be made by the sample of therapists providing reports on families. The suitability of these values depends on the intended use of the instrument, with higher values required for higher stakes decision making. For the Caregiver dimension, the person separation reliability at T4 was 3.97 (Person Mean *S.E.* = .23) and for the items it was 4.18 (Person Mean *S.E.* = .24). For the Youth dimension, the person separation reliability was 3.79 (Person Mean *S.E.* = .34) and for the items it was 5.35 (Item Mean *S.E.* = .48). In this case, the ability to make four or more distinctions in the sample of items or persons, reflecting approximately five distinct levels, is judged to be entirely sufficient. For

[5]Rasch item analyses ignored the multilevel nature of the data. As a check on whether this affected substantive conclusions, Rasch analyses of item properties were rerun with a sample of 42 families, with one family randomly selected for each therapist. The Rasch analyses results from this sample were compared to the non-independent data results. Negligible differences in the fit statistics and *SEs* were observed in these comparisons, suggesting the actual role of dependence in the complete data set had little impact on the Rasch analyses.

the T3 data, with one exception, the results were the same. The person reliability estimates for the Youth dimension were slightly lower, with values of .87 and 2.59 for reliability and separation reliability, respectively.[6]

Rasch Item-Person Map

Rasch Item-Person Maps provide visual displays of the extent to which family scores represent a full range of values. They also indicate where each item best differentiates treatment outcome among families – whether it is a "hard" item (only endorsed by therapists for highly successful families) or an "easy" item (endorsed by therapists for families who have made some progress but who may not have been completely successful). Figure 1 summarizes the results of the T4 Rasch rating scale models for the Caregiver dimension. On the left, in the Person column, the symbols reflect the distribution of families. This distribution is ordered such that families at the top were judged by the therapist to have better treatment outcomes, families in the middle are at the mean level, and families at the bottom were judged to have worse treatment outcomes. Overall, the family distribution has a wide range, reflecting significant variability in outcomes as judged by the sample of therapists. Importantly, the estimated locations of the families are reasonably precise, as reflected by the reliability results reported above.

The next three columns depict the distribution of items and the targeting of the items to the sample of families. Ideally, the distribution of items will cover the full distribution of families. The range of item measures is relatively narrow, covering a span of approximately 2.5 logits. Despite this, most of the distribution of families is targeted by the items (likely due to the use of a 5-point rating scale). The item column on the left reflects the portion of the sample covered by the items when there is a 50% probability of being rated in the lowest rating scale category. The middle column reflects the location of the mean item measures (i.e., equal probability of being rated in the lowest and highest categories), and the column on the right reflects the range of the sample covered by the items when there is a 50% probability of being rated in the highest category (or lower). Thus, the rating scale provides reasonable coverage of the distribution of people; however, gaps exist around the higher levels of person measures.

Figure 2 similarly summarizes the T4 Rasch rating scale models for the Youth dimension. Based on the Person distribution, therapist judgments of outcomes are broadly distributed, suggesting significant variability in outcomes. Similar to the Caregiver dimension, the estimated locations of the families are also reasonably precise. The range of item distribution is somewhat narrow (approximately 2.0 logits) and not evenly distributed,

[6]Another way of evaluating item consistency with a new instrument is to examine item invariance, or whether items show similar difficulty levels across administrations or samples. To evaluate item invariance (i.e., the stability of estimated item difficulty between T3 and T4), we performed a simultaneous calibration, common person equating approach with each item specified as two different versions – a T3 version and a T4 version. This model estimates the measure of each item at each occasion on the same scale of measurement. Using methods to identify significant differences, developed by Wright and Stone (1979) and Bond and Fox (2007), we (a) cross-plotted the T3 and T4 item measures, (b) plotted an identity reference line, and (c) used the *SE* estimates for each item to compute and plot 95% control lines around the scatter of item measures. Only one item (Item 3; The family has met the overarching goals of treatment) was plotted outside the 95% confidence region, suggesting significantly different item performance between T3 and T4. Item content suggests this item in particular is most reflective of successful treatment termination and would be expected to change between mid and post-treatment. Thus the results supported the assumption that items perform similarly across time and that Youth and Caregiver scores will reflect treatment outcome in similar ways at mid and post-treatment.

suggesting some gaps in coverage by items around high levels of youth outcome behaviors, in particular with favorable outcomes.

Reexamination of Dimensionality

To check whether dimensionality still was present in the data after collapsing the rating scale and eliminating misfitting items, a second Rasch PCA of the TPTO:YAB items was performed using the 5-point rating scale and excluding misfitting items. Again this indicated two dimensions at T4 with the expected items patterns and meaningful dimensionality (67.0% of the variance explained in the first constrast; eigenvalue of 5.2).[7] As previously, dimensionality assessments conducted separately for Caregiver and Youth dimensions showed no further meaningful dimensionality. At T4, 73.8% (Caregiver) and 83.3% (Youth) of the total variance in the dimension scores was explained by Rasch item and person measures. The eigenvalues of the first contrast for both dimensions (1.9 for Caregiver dimension; 1.5 for Youth dimension) suggested there was not meaningful dimensionality (Linacre, 2013). Item loading data supported this conclusion as did identical analyses at T3.

These dimensionality analyses suggested the TPTO:YAB had two separate subscales: an Effective Caregiver Outcome (ECO) subscale and a Youth Behavior Outcome (YBO) subscales. Scores for these subscales correlated highly at both time points, r =.47 at T3 and .62 at T4, $ps < .001$, however, suggesting that they might share significant variance and perhaps reflect a higher order construct.

Bifactor models (Reise, 2012) test whether an instrument consists of a broader construct that is assessed with items related to identifiable “subdomains” – in this case, the extent to which the caregiver and youth dimensions were independent from the broader underlying construct of how the therapist generally perceived the family's progress in therapy. To examine this issue and to inform scoring and interpretive decisions about the TPTO:YAB, we used Reise's (2012) procedures to examine whether a confirmatory bifactor model might be appropriately applied to TPTO:YAB items. The models were performed using IRTPRO software (v2.1; Cai, Thissen, & du Toit, 2011). Our 14 final, retained items were supplied, with each item specified loading on the general dimension, five items loading on the Youth dimension, and nine items loading on the Caregiver dimension. These analyses also provided McDonald's omegas (ω), statistics that provide estimates of the reliability of the specific (i.e., Youth, Caregiver) dimensions before and after removing the effect of their loadings on the general factor (ω_H, ω_S). McDonald's ω is also used to estimate the reliability of the general dimension.

Table 4 provides results of the analyses. At T4, for the general dimension, the item loadings were high, notably: 10 items loaded at .80 or above, 2 loaded at .79 and .73, 2 about .50. On the specific dimensions, the loadings were lower, though non-trivial. For the Youth dimension, the loadings ranged from .31-.59, and for the Caregiver dimension, .27-.70. Reliability statistics computed based on these results indicated McDonald's ω = .83 for the index of reliability based on the percentage of common variance attributable to the general

[7]A multilevel factor analysis with the retained items and 5-point scoring also supported these findings; specifics available upon request.

factor. Model-based reliability was also high, ω_H = .97 (ω_H = .94 for Youth, ω_H = .95 for Caregiver dimensions individually). Removing the effects of the general dimension, these values decrease substantially. For the Youth dimension, ω_S = .24 and for the Caregiver dimension, ω_S = .22. Combined these results suggest that, despite evidence for a bifactor structure and dimensionality, the individual Youth and Caregiver dimensions largely reflect the single source of variance from the general dimension.

The same model was tested using the T3 data, and the results were generally consistent (see Table 4). The loadings on the general dimension remained high. Loadings on the Youth dimension, however, were stronger than at T4, whereas loadings on the Caregiver dimension were generally weaker. For the Youth dimension, ω_S = .49, higher than with the T4 data, though still a low level of reliability. For the Caregiver dimension, ω_S = .12, a decrease from the T4 data.

At both time points, these results indicate very limited unique variance contributed by the Caregiver and Youth dimensions. Although the two separate dimensions were clearly identified, each provides modest information about the specific effective caregiver or youth behavior outcomes but instead largely reflect the broader domain of treatment outcome.

TPTO:YAB Scoring

To evaluate validity evidence for the TPTO:YAB, raw mean scores were computed for analysis with items and dimensions retained and defined based on Rasch and bifactor analysis results. The dimensionality analyses suggested the presence of two dimensions: the Effective Caregiver Outcome scale (ECO; Items 2, 4, 7, 9, 12, 13, 14, 17, and 20) and Youth Behavior Outcome scale (YBO; Items 1, 3, 5, 8, 10); see Table 1 for items. Mean scores from these two scales were used for the remaining validity evaluations.[8] Scores were based on the 5-point rating scale and items 2, 3, 4, 5, 7, 10, 12, 13, 14, 17, and 20 were reverse-scored so that high scores suggest favorable treatment outcomes.

The bifactor analysis results suggested scores from the two scales suffer from a lack of uniqueness. In light of these findings, and because an omnibus indicator assessing therapist view of outcome might be preferred over subscale scores in certain settings (e.g., in research contexts where numbers of analyses need to be minimized), a total TPTO:YAB score was also included in these analyses, in addition to the ECO and YBO subscale scores. This total score was the mean of the 14 items that comprise the ECO and YBO scales. At T3 and T4, coefficient alphas for the ECO scale were .95 and .96, and .92 and .96 for the YBO scale. Coefficient alpha for the total scale was .95 at T3 and .96 at T4. The minimum bivariate *r* between pairs of items was .28 and .30 at T3 and T4, respectively; average *r*s were .56 (T3) and .65 (T4). Table 5 presents means and standard deviations for scores from the TPTO:YAB and other measurement tools used in the study.

[8]A benefit of Rasch analyses is the ability to use Rasch logit measures (instead of raw scores) because these measures reduce off-construct “noise” in the scores. Nonetheless, unweighted item mean scores are used in the validity analyses because: (a) mean scores are often more practical and the utility of the measure will be better assessed if the analyses are conducted using scores that are generalizable to applied settings; (b) the ECO and YBO mean scores correlated very highly with Rasch measures of the same dimensions (ECO r = .98, n = 162, p < .001; YBO r = .97, n = 163, p < .001 at Time 4; ECO r = 1.00, n = 111, p < .001; YBO r = .99, n = 111, p < .001 at Time 3), suggesting analysis results would be nearly identical. However, the first author can provide a logit conversion table upon request.

Temporal Consistency

TPTO:YAB score temporal consistency was assessed from T3 to T4 using multilevel correlations for total, ECO, and YOB scores (total and ECO n = 100; YOB n = 101). On average, T4 TPTO:YAB assessments occurred 11 weeks after T3 assessments (SD = 6 weeks, range 3-27 weeks). Moderate correlations were expected based on the assumption that families responding well to therapy in mid-treatment would for the most part also be responding well at the end of treatment. TPTO:YAB total scores at T3 (M = 3.28, range = 3.71) correlated strongly with TPTO:YAB total scores at T4 (M = 3.50, range = 3.93), r = .78, p < .001, as did ECO scale scores (T3 M = 3.27, range = 3.78; T4 M = 3.44, range = 4.00), r = .81, p < .001. TPTO:YAB YBO scale scores correlated moderately between T3 (M = 3.29, range = 4.00) and T4 (M = 3.60, range = 4.00), r = .58, p < .001. These results provide preliminary evidence for TPTO:YAB score reliability across time, given that treatment was ongoing from T3 to T4 and some individual differences in shifts in outcomes – particularly in youth outcome behaviors, the ultimate goals of treatment – would be expected. These reliability coefficients only assess temporal consistency, not inter-therapist agreement. Correlations may be inflated because the same informant completed the instrument at T3 and T4.

Validity Evidence based on Associations with Other Variables

We expected to find evidence for the valid use of TPTO:YAB scores based on correlations between each scale scores of the TPTO:YAB and scores from five tools assessing related or similar constructs both at T3 and T4 (the APQ Inconsistent Discipline, APQ Poor Monitoring, SRD General Delinquency, SIPA Incompetence/Guilt, a CBCL Externalizing scale scores; see Tables 6 and 7). We calculated these correlations two ways. The first used MPlus to model correlations taking into account the multilevel nature of the data (see Table 6), and may be most generalizable to research uses in which therapists rate more than one client and this nonindependence is considered in statistical analyses. Because these correlations control for between therapist differences, we supplemented these using SPSS to calculate bivariate (disaggregated) correlations that ignored nonindependence of therapist ratings (see Table 7). The multilevel correlations with scores assessing related constructs provided evidence that TPTO:YAB total, ECO, and YBO scores at T4 as well as the Total and YBO scores at T3 provided information relevant to treatment outcome. Correlations involving the ECO scale scores were weaker at T3, however. In addition, multilevel correlations were generally higher than disaggregated correlations that failed to control for therapist differences, although both sets of correlations showed the same pattern of relationships. This discrepancy may have resulted from the fact that TPTO:YAB scores were more similar across cases seen by the same therapist than were scores provided by caregivers and youths. Specifically, ICCs for TPTO:YAB scores ranged from .20-.34, while ICCS for the other measurement tools ranged from .01-.06.

We expected T3 and T4 TPTO:YAB scores to be associated with T3 and T4 CBCL internalizing scores but with weaker magnitude than T3 and T4 CBCL externalizing scores (see Tables 6 and 7). No statistically significant differences were found between TPTO:YAB score correlations with CBCL internalizing and externalizing scores using tests of differences between dependent correlations (Bruning & Kintz, 1987). However, further

assessment of the CBCL scales revealed particularly high correlations between the Internalizing and Externalizing scales at T3, $r = .75$, $p < .001$, $n = 103$, and T4, $r = .76$, $p < .001$, $n = 168$. These correlations suggest the Internalizing scale scores of the CBCL may not have provided information that was clearly distinct from the Externalizing scale scores in this sample.

We also examined whether T4 TPTO:YAB scores differentiated between families identified as terminating successfully ($n = 62$ at T4) versus non-successfully ($n = 100$ at T4), based on CDS data. Multilevel logistic regression analyses (using HLM v. 6.08; Raudenbush, Bryk, & Congdon, 2009) with CDS treatment success as the outcome and T4 TPTO:YAB scores as predictors indicated TPTO:YAB scores were significantly better for successful than for unsuccessful terminators. Specifically, for a one-point increase in TPTO:YAB average total scores at T4, participants were 7.60 times more likely to be in the successful termination group at T4, $b = 2.03$, S.E. = .39, $z = 5.21$, $p \leq .001$, $n = 162$. For a one-point increase in the average score of the ECO and YO TPTO:YAB scales at T4, participants were 6.33, $b = 1.85$, S.E. = .37, $z = 5.01$, $p \leq .001$, $n = 162$, and 3.75, $b = 1.32$, S.E. = .26, $z = 5.18$, $p \leq .001$, $n = 163$, times more likely to be in the successful termination group at T4, respectively.

Finally, we examined whether mid-treatment (T3) TPTO:YAB scores predicted successful treatment termination. Specifically, multilevel regression analyses (controlling for therapist level effects) were used to examine whether each of the T3 TPTO:YAB scores predicted T4 CBCL externalizing scores and CDS termination success. Because SRD scores showed minimal effects of therapist differences (ICC=.01) and involved count data, negative binomial analyses were used to predict this variable. None of the T3 TPTO:YAB scores significantly predicted T4 SRD scores. Time 3 TPTO:YAB YBO scale scores significantly predicted CBCL scores at T4, $b = -.36$, S.E. = .17, $t = -2.07$, $p \leq .05$, $n = 96$. T3 TPTO:YAB total and ECO scale scores did not significantly predict T4 CBCL scores. In contrast, all three T3 TPTO:YAB scores significantly predicted successful termination decisions at T4. An increase of 1 point in average T3 TPTO:YAB total, ECO, and YBO scores suggested participants were 4.96, $b = 1.60$, *S.E.* = .44, $z = 3.61$, $p \leq .001$, $n = 100$; 3.35, $b = 1.21$, *S.E.* = .35, $z = 3.48$, $p = .001$, $n = 100$; and 3.38, $b = 1.22$, *S.E.* = .38, $z = 3.25$, $p = .001$, $n = 100$, times more likely to be in the successful termination group, respectively. The majority of positive findings are based on associations between TPTO:YAB scores and scores from another therapist-completed measure, therefore, shared method variance should be considered in the interpretation of these findings.

Discussion

The present study provided an initial investigation of the psychometric properties of the TPTO:YAB, a therapist-completed measure of outcome designed to be used in evidence-based treatment for youth antisocial behavior. Four central findings of this study should be highlighted. First, we found evidence of both differentiated and global therapist views of family functioning late in treatment. Of the original 20 items of the TPTO:YAB, 14 items could be extracted to compose two dimensions, one examining Effective Caregiver Outcomes and the other describing Youth Behavior Outcomes, which were distinct both in mid-treatment and at the end of treatment. The identification of these two dimensions

supports previous empirical links of parental factors in the development and maintenance of antisocial behavior in youth (McMahon, Wells, & Kotler, 2006) and is consistent with the MST theory of change, which stipulates that caregiver change is essential for youth behavior change (Henggeler et al., 1998).

At the same time, bifactor analysis results raised questions about the level of precision measured by the ECO and YBO scales beyond the overall construct of treatment outcome, and suggested that ratings on these items are greatly influenced by how the family is progressing more generally. Perhaps therapists form a general impression or conclusion about how families are responding in MST, and this informs how they rate both caregivers and youth, despite some differentiation in how they rate the different participants in treatment. It may also be the case that for the most part caregiver and youth changes proceed in tandem, as suggested by the MST theory of change. By the end of treatment these should be in sync, for the most part, so an underlying common dimension would be expected because of the correlated change that happens when treatment is successful.

Second, subscale and total TPTO:YAB scores showed preliminary evidence of good internal reliability. In addition, Rasch item and rating scale analyses supported the reliability of ECO and YBO scales. Rasch analyses and person-item maps also indicated that caregiver and youth scores reflect individual differences among families, supporting their use to describe outcomes in research and clinical settings. TPTO:YAB scores correlated from mid-treatment to termination at expected levels, given that treatment was ongoing. It is important to note that the temporal consistency results may be inflated due to the brief time interval between time periods, the potential for therapist memory effects, or the potential for therapists to resist changing their impressions of clients (Garb, 1998, 2004). A more rigorous study of TPTO:YAB score temporal consistency would be needed to strengthen these findings. This could include comparing scores with longer intervals between ratings or comparing therapists scores at one time point with scores from a co-therapist or clinically-trained session observer with full access to information about the family at a later time point (Sultan et al., 2006), ideally using a generalizability study.

Third, validity analyses generally supported the use of the total score as an indicator of family outcome. Total scores showed significant, mostly moderate relationships with most caregiver-rated measures of parenting practices and youth- and caregiver-rated measures of externalizing behaviors at treatment termination, with somewhat weaker correlations in mid-treatment. In addition, the YBO scale scores correlated significantly albeit modestly with CBCL externalizing behavior scores at mid-treatment (r = .33). At treatment termination, this correlation was substantially higher (r = .52), and YBO scores also were associated with self-report delinquency scores.

In contrast, correlations with parent-rated measures of positive and negative parenting behaviors provided limited validity evidence for the ECO scale scores at either time point. This could have been due in part to the different foci of the measures. The ECO scale items measured a variety of fairly global aspects of caregiving (e.g., problem-solving, doing what is necessary for the youth to succeed); APQ and SIPA items assessed more specific aspects of parenting and perceived competence. YBO scores, which mapped more closely onto the

symptom counts assessed by the SRD and CBCL, correlated more highly with scores from these validation measures.

YBO, ECO, and total scores at treatment termination were significantly higher for those who terminated based on mutual decisions that the family had met treatment goals versus those who terminated under other circumstances. Exploratory predictive analyses also indicated that mid-treatment TPTO:YAB total, ECO, and YBO scores predicted later termination success. However, termination success was determined based on data from a tool completed by therapists (CDS). With therapists completing both measures, we expect some shared method variance to inflate correlations between these scores, however the CDS is a report of termination circumstances and presumably based on factual information, not focused on therapist judgments of outcome factors. Nonetheless, further investigation of the relationship of TPTO:YAB scores with other omnibus indicators of success provided by different informants would be useful.

Finally, the results provide insights into how therapists evaluate treatment outcome, at least when asked to make judgments on rating scales like the TPTO:YAB. As noted above, findings from the bifactor analysis indicated that therapists may make a general judgment about how the family has responded to treatment, and this broad judgment is reflected in their ratings of specific dimensions. The TPTO:YAB was designed to capture therapist general judgment about outcome, so this may have been a by-product of the ways the scale was created. Also, intraclass correlations in this study suggested that therapist differences accounted for nontrivial variance in TPTO:YAB scores. Although this could be a function of some therapists being more skillful (or seeing more responsive families) than others, ICCs for other outcome measures (e.g., SRD, CBCL) were much lower, suggesting therapists provide ratings in a somewhat systematic manner across their cases. Possibly as a result, validity correlations were somewhat attenuated with analyses that failed to account for these within-therapist relationships. The contribution of informant differences to scores involving ratings by others has been frequently noted and discussed in various settings (e.g., Garb, 1998; Hoyt & Kerns, 1999). Hoyt and Kerns (1999) in particular distinguish between rater variance (differences due to individual therapists systematically rating clients more severely or more leniently) and dyadic variance (more idiosyncratic differences in ratings due to some therapists rating some clients differently, e.g., rating individual families more highly if they like them). Multilevel analyses can address the first of these and will be possible in research contexts in which multiple therapists are involved and each sees more than one client. Other methods of estimating and correcting for bias are available as well (Hoyt, 2000), although many of these will not be practical in many research and clinical contexts (Hoyt, 2002). Nonetheless, understanding contributors to variance in therapist ratings will be important to consider, especially in applied decision-making using therapist reports of outcome.

Item Content and Treatment Outcome

Rasch results also indicated that therapists endorsed some items on each subscale more readily than others. This information supports the validity of the results, and also provides substantive information about which aspects of outcomes are more difficult to attain than

others. For the youth scale, “The youth's problem behavior has improved,” was the most readily endorsed item, whereas “The youth's behavior places him/her at risk for arrest or placement outside the home” was the least. This order is logical: improvement in behavior does not guarantee that the youth is has changed substantially or consistently. For the Caregiver scale, the most difficult items were, “The caregivers’ own life issues keep them from parenting effectively, the caregivers consistently use appropriate parenting practices,” and “the caregivers can deal with the youth effectively without needing my (the therapist's) advice,” whereas the most readily endorsed was, “the caregivers will do whatever is necessary to help the youth succeed.” This suggests that engaging and motivating caregivers may be more probable than directly propelling consistent changes in the behaviors exhibited by youth with antisocial behavior problems. In addition, these findings point to the persistent issue of helping caregivers parent effectively even when their own difficulties interfere.

These analyses also indicated that the ranges of items for the ECO and YBO scales were somewhat limited, particularly in the high and low ends of treatment outcome (especially the YBO scale). The scales may be improved by adding more difficult items (i.e., items that are only strongly endorsed for particularly successful families), such as items assessing therapist judgment of low future risk and sustained, consistent, and substantial behavior change.

Use of the TPTO:YAB

As noted by Garland et al. (2003), “Despite all of the research, administrative, and policy attention to outcome measurement in mental health services, the actual clinical utility of outcome measurement remains largely unexamined” (p. 393). As developers and providers of ESTs make efforts to increase provider accountability and improve efficient, focused assessment of outcomes, there is a need for practical instruments that can be used to make reliable and valid decisions about treatment outcome (Ebesutani et al., 2012). An instrument that captures therapist judgments regarding treatment outcome can help address common barriers of outcome measurement (i.e., feasibility of administration), while providing data from an often unutilized respondent. In this study, the results generally suggest that the TPTO:YAB shows promise as a therapist completed tool that can supplement other indicators of treatment outcome for families of antisocial youth receiving MST or related ESTs. In research settings, the TPTO:YAB provides an additional tool for assessing outcome in studies of family therapy with antisocial youth. The TPTO:YAB could also be useful when data from families are difficult or impossible to collect.

TPTO:YAB items were developed with treatment outcome in mind. That is to say, end of treatment evaluation was intended, rather than mid-treatment evaluation. Nonetheless, we examined both time points and found that item and scale performance was generally consistent across time points, albeit with weaker validity correlations at mid-treatment. Possibly TPTO:YAB scores could have significant impact on mid-treatment corrections, tritrated according to youth/family responsiveness, but further research would be needed to support this use.

In addition, although ECO and YBO items are face valid for measuring treatment factors connected with either effective caregiver outcomes or youth behavior outcomes, bifactor analyses of the specific reliability of these scores indicated that the influence of caregiver

and youth outcome behaviors may be overshadowed by the overall perception of treatment outcome provided by therapists. Therefore, these scales will not be appropriate if the goal of the assessment is to examine youth- or caregiver-specific outcomes, independent of variance contributed by the therapist's overall perception of family response to treatment. Additional items should be created and evaluated as augmentation to the ECO and YBO scales to achieve these goals. At this point, the total TPTO:YAB score has the strongest evidence that it can be used to provide a general description of therapist evaluation of family response to MST.

Limitations and Future Directions

We did not find evidence that TPTO:YAB scores showed differential relationships with child internalizing and externalizing symptomology at termination. One explanation for this is that therapists perceive internalizing and externalizing behaviors as equally relevant in judging treatment outcome. Alternatively, the finding could be due to imprecise measurement of internalizing behavior. CBCL Externalizing and Internalizing scores correlated very highly at mid- and post-treatment, suggesting that the Internalizing CBCL scores did not provide information that was distinct from the Externalizing scores in this sample. Others have also noted this overlap, especially among clinical samples (e.g., Seligman et al., 2004; Stanger & Lewis, 1993). Future studies evaluating the differential relationship of TPTO:YAB scores to supposedly distinct constructs should select alternative instruments that measure constructs less strongly related to antisocial behavior.

Because this study is the first to provide data on the TPTO:YAB, replications are needed with different samples. Further examination of the utility of the TPTO:YAB could explore cultural considerations with groups that were not included in the current study. Additional research could also examine the use of the TPTO:YAB in broader contexts, such as families participating in other treatment approaches.

Careful consideration of contextual therapist factors that could influence response bias would be important to examine in future research. An investigation that explicitly examined therapist agreement in TPTO:YAB (perhaps between cotherapists seeing the same family) would be useful. It is also likely that the settings in which therapist perspectives will be most valuable will exclude situations in which scores will influence therapist compensation or performance evaluations, as they may be prone to response biases based on social desirability or leniency biases, depending on the purpose of the scores (Podsakoff & MacKenzie, 2003). Further exploration of therapist biases in rating families will also have important implications for use of the instrument in applied settings. Ultimately, further evaluation of the TPTO:YAB could contribute to identifying those circumstances in which practitioner perspectives can be utilized to provide valid and reliable information.

Acknowledgments

The authors would like to thank the research team involved in data collection for this study, the staff, and families receiving treatment at the University of Colorado Denver Synergy Program, the University of Colorado Hospital Outpatient Community-Based Services, Savio House, and Jefferson County Mental Health, as well as Angi Wold, Irwin S. Rosenfarb, and Fernando A. Ortiz.

This study was supported in part by National Institute of Mental Health funding: R01MH68813 (Phillippe B. Cunningham, P.I.). The fourth author is a paid consultant of MST Services and is part owner of Evidence Based Services, Inc., a MST Network Partner Organization. The TPTO:YAB may be used without cost by researchers and clinicians; please credit this article as the original source.

References

Achenbach, TM. Manual for the Child Behavior Checklist/4-18 and 1991 Profile. University of Vermont Department of Psychiatry; Burlington, VT: 1991.

Bond, TG.; Fox, CM. Applying the Rasch Model. Lawrence Erlbaum Associates, Publishers; Mahwah, New Jersey: 2007.

Bruning, JL.; Kintz, BL. Computational handbook of statistics. 3rd ed.. Harper Collins; Glenview, IL: 1987.

Cai, L.; Thissen, D.; du Toit, SHC. IRTPRO for Windows [Computer software & manual]. Scientific Software International; Lincolnwood, IL: 2011.

Clark LA, Watson D. Constructing validity: Basic issues in objective scale development. Psychological Assessment. 1995; 7:309–319. doi:10.1037//1040-3590.7.3.309.

Cohen, J.; Cohen, P.; West, SG.; Aiken, LS. Applied multiple regression/correlation analysis for the behavioral sciences. 3rd ed.. Erlbaum; Hillsdale: 2003.

Dunn TW, Burlingame GM, Walbridge M, Smith J, Crum MJ. Outcome assessment for children and adolescents: Psychometric validation of the Youth Outcome Questionnaire 30.1 (Y-OQ-30.1). Clinical Psychology and Psychotherapy. 2005; 12:388–401. doi:10.1002/cpp.461.

Dutra L, Campbell L, Westen D. Quantifying clinical judgment in assessment of adolescent psychopathology: Reliability, validity, and factor structure of the Child Behavior Checklist for clinician report. Journal of Clinical Psychology. 2004; 60:65–85. doi:10.1002/jclp.10234. [PubMed: 14692010]

Ebesutani C, Bernstein A, Chorpita BF, Weisz JR. A transportable assessment protocol for prescribing youth psychosocial treatments in real-world settings: Reducing assessment burden via self-report scales. Psychological Assessment. 2012; 24:141. doi: 10.1037/a0025176. [PubMed: 21859220]

Elliott, DS.; Ageton, SS.; Huizinga, D.; Knowles, BA.; Canter, RJ. The prevalence and incidence of delinquent behavior: 1976–1980 (National Youth Survey Report No. 26). Behavioral Research Institute; Boulder, CO: 1983.

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

Frick, PJ. The Alabama Parenting Questionnaire. University of Alabama; 1991. Unpublished instrument

Froyd JE, Lambert MJ, Froyd JD. A review of practices of psychotherapy outcome measurement. Journal of Mental Health. 1996; 5:11–15. doi:10.1080/09638239650037144.

Garb, HN. Studying the clinician: Judgment research and psychological assessment. APA Books; Washington, DC: 1998.

Garb HN. Clinical judgment and decision making. Annual review of clinical psychology. 2004; 1:67–89. doi: 10.1146/annurev.clinpsy.1.102803.143810.

Garland AF, Kruse M, Aarons GA. Clinicians and outcome measurement: What's the use? Journal of Behavioral Health Services & Research. 2003; 30:393–405. doi:10.1007/BF02287427. [PubMed: 14593663]

Hawes DJ, Dadds MR. Assessing parenting practices through parent-report and direct observation during parent-training. Journal of Child and Family Studies. 2006; 15:555–568. doi:10.1007/s10826-006-9029-x.

Haynes SN, Richard DCS, Kubany ES. Content validity in psychological assessment: A functional approach to concepts and methods. Psychological Assessment. 1995; 7:238–247. doi: 10.1037/1040-3590.7.3.238.

Henggeler, SW. Delinquency in adolescence. Sage; Newbury Park, CA: 1989.

Henggeler SW, Rodick JD, Borduin CM, Hanson CL, Watson SM, Urey JR. Multisystemic treatment of juvenile offenders: Effects on adolescent behavior and family interactions. Developmental Psychology. 1986; 22:132–141.

Henggeler SW, Schaeffer CM. Treating serious emotional and behavioural problems using Multisystemic Therapy. The Australian and New Zealand Journal of Family Therapy. 2010; 31:149–164. doi:10.1375/anft.31.2.149.

Henggeler, SW.; Schoenwald, SK.; Borduin, CM.; Rowland, MD.; Cunningham, PB. Multisystemic Treatment of antisocial behavior in children and adolescents. The Guilford Press; New York: 1998.

Henggeler, SW.; Schoenwald, SK.; Borduin, CM.; Rowland, MD.; Cunningham, PB. Mulitsystemic Treatment of antisocial behavior in children and adolescents. 2nd ed.. The Guilford Press; New York: 2009.

Henggeler, SW.; Schoenwald, SK.; Rowland, MD.; Cunningham, PB. Serious emotional disturbances in children and adolescents. The Guilford Press; New York: 2002.

Hoge, RD.; Andrews, DA. The Youth Level of Service/Case Management Inventory manual and scoring key. Multi-Health Systems; Toronto, Canada: 2002.

Hoyt WT. Rater bias in psychological research: When is it a problem and what can we do about it? Psychological Methods. 2000; 5:64–86. doi: 10.1037/1082-989X.5.1.64. [PubMed: 10937323]

Hoyt WT. Bias in participant ratings of psychotherapy process: An initial generalizability study. Journal of Counseling Psychology. 2002; 49:35–46. doi: 10.1037/0022-0167.49.1.35.

Hoyt WT, Kerns MD. Magnitude and moderators of bias in observer ratings: A meta-analysis. Psychological Methods. 1999; 4:403–424. doi: 10.1037/1082-989X.4.4.403.

Huey SJ, Henggeler SW, Brondino MJ, Pickrel SG. Mechanisms of change in Multisystemic Therapy: Reducing delinquent behavior through therapist adherence and improved family and peer functioning. Journal of Consulting and Clinical Psychology. 2000; 68:452–467. doi: 10.1037//0022-006X.68.3.451.

Huizinga D, Elliott DS. Reassessing the reliability and validity of self-delinquent measures. Journal of Quantitative Criminology. 1986; 2:293–327. doi: 10.1007/BF01064258.

Kazak AE, Hoagwood K, Weisz JR, Hood K, Kratochwill TR, Vargas LA, Banez GA. A meta-systems approach to evidence-based practice for children and adolescents. American Psychologist. 2010; 65:85–97. doi:10.1037/a0017784. [PubMed: 20141264]

Kazdin AE. Nonspecific treatment factors in psychotherapy outcome research. Journal of Consulting and Clinical Psychology. 1979; 47:846–852. doi:10.1037//0022-006X.47.5.846. [PubMed: 512143]

Kazdin AE, Wassel G. Treatment completion and therapeutic change among children referred for outpatient therapy. Professional Psychology: Research and Practice. 1998; 29:332–340. doi: 10.1037//0735-7028.29.4.332.

Lau AS, Garland AF, Yeh M, McCabe KM, Wood PA, Hough RL. Race/ethnicity and inter-informant agreement in assessing adolescent psychopathology. Journal of Emotional and Behavioral Disorders. 2004; 12:145–156. doi:10.1177/10634266040120030201.

Leibert T. Making change visible: The possibilities in assessing mental health counseling outcomes. Journal of Counseling and Development. 2006; 84:108–113. doi:10.1002/j.1556-6678.2006.tb00384.x.

Linacre JM. Optimizing rating scale category effectiveness. Journal of Applied Measurement. 2002; 3:85–106. [PubMed: 11997586]

Linacre, JM. Winsteps computer program for many-facet Rasch measurement (version 3.80.1) [Computer software and manual]. Winsteps.com; Beaverton, Oregon: 2013.

McMahon, RJ.; Frick, PJ. Conduct and oppositional disorders.. In: Mash, EJ.; Barkley, RA., editors. Assessment of childhood disorders. 4th ed.. Guilford Press; New York: 2007. p. 132-183.

McMahon, RJ.; Wells, KC.; Kotler, JS. Conduct problems.. In: Mash, EJ.; Barkley, RA., editors. Treatment of childhood disorders. 3rd ed.. Guilford Press; New York: 2006. p. 137-268.

Meyer GJ, Finn SE, Eyde LD, Kay GG, Moreland KL, Dies RR, Reed GM. Psychological testing and psychological assessment: A review of evidence and issues. American Psychologist. 2001; 56:128–165. doi:10.1037/0003-066X.56.2.128. [PubMed: 11279806]

Morey, LC. Personality Assessment Screener professional manual. Psychological Assessment Resources; Odessa, FL: 1997.

Murphy MJ, Faulkner RA, Behrens C. The effect of therapist-client racial similarity on client satisfaction and therapist evaluation of treatment. Contemporary Family Therapy: An International Journal. 2004; 26:279–292. doi: 10.1023/B:COFT.0000037915.95303.28.

Muthén, LK.; Muthén, BO. Mplus user's guide. 6th ed.. Muthén & Muthén; Los Angeles, CA: 1998-2010.

Podsakoff PM, MacKenszie SB, Lee Jeong-Yeon, Podsakoff NP. Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology. 2003; 88:879–903. doi:10.1037/0021-9010.88.5.879. [PubMed: 14516251]

Raudenbush, SW.; Bryk, AS.; Congdon, R. *HLM 6: Hierarchical linear & nonlinear modeling* (version 6.08) [Computer software & manual].. Scientific Software International; Lincolnwood, IL: 2009.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Dankmarks Paedagogiske Institut; Copenhagen, Denmark: 1960.

Rasch, G. Probabilistic models for some intelligence and attainment tests. expanded ed.. University of Chicago Press; Chicago: 1980.

Reise SP. The rediscovery of bifactor measurement models. Multivariate Behavioral Research. 2012; 47:667–696. [PubMed: 24049214]

Reitman D, Hummel R, Franz DZ, Gross AM. A review of methods and instruments for assessing externalizing disorders: theoretical and practical considerations in rendering a diagnosis. Clinical Psychology Review. 1998; 18:555–584. [PubMed: 9740978]

Rosen A, Proctor EK. Distinctions between treatment outcomes and their implications for treatment evaluation. Journal of Consulting and Clinical Psychology. 1981; 49:418–425. doi: 10.1037//0022-006X.49.3.418. [PubMed: 7276331]

Ruscio J, Roche B. Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. Psychological Assessment. 2012; 24:282–292. doi: 10.1037/a0025697. [PubMed: 21966933]

Salekin, RT.; Salekin, KL.; Clements, CB.; Leistigo, AR. Risk-Sophistication-Treatment Inventory.. In: Grisso, T.; Vincent, G.; Seagrave, D., editors. Mental health screening and assessment in juvenile justice. Guilford Press; New York: 2005. p. 341-356.

Sawyer MG, Sarris A, Baghurst PA, Cornish CA, et al. The prevalence of emotional and behaviour disorders and patterns of service utilization in children and adolescent. Australian and New Zealand Journal of Psychiatry. 1990; 24:323–330. doi:10.3109/00048679009077699. [PubMed: 2241716]

Schoenwald SK, Sheidow AJ, Letourneau EJ, Liao JG. Transportability of Multisystemic Therapy: Evidence for multilevel influences. Mental Health Services Research. 2003; 5:223–239. doi: 10.1023/A:1026229102151. [PubMed: 14672501]

Seligman LD, Ollendick TH, Langley AK, Baldacci HB. The utility of measures of child and adolescent anxiety: a meta-analytic review of the Revised Children's Manifest Anxiety Scale, the State–Trait Anxiety Inventory for Children, and the Child Behavior Checklist. Journal of Clinical Child and Adolescent Psychology. 2004; 33:557–565. [PubMed: 15271613]

Sheras, PL.; Abidin, RR.; Konold, TR. Stress Index for Parents of Adolescents (SIPA). Psychological Assessment Resources; Odessa, FL: 1998.

Smith EV Jr. Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. Journal of Applied Measurement. 2001; 2:281–311. [PubMed: 12011511]

Smith EV Jr. Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement. 2002; 3:205–231. [PubMed: 12011501]

Smith RM. Fit analysis in latent trait measurement models. Journal of Applied Measurement. 2000; 1:199–218. [PubMed: 12029178]

Smith RM, Schumacker RE, Bush MJ. Using item mean squares to evaluate fit to the Rasch model. Journal of Outcome Measurement. 1997; 2:66–78. [PubMed: 9661732]

Stanger C, Lewis M. Agreement among parents, teachers, and children on internalizing and externalizing behavior problems. Journal of Clinical Child Psychology. 1993; 22:107–116.

Sultan S, Andronikof A, Réveillère C, Lemmel G. A Rorschach stability study in a nonpatient adult sample. Journal of Personality Assessment. 2006; 87:330–348. [PubMed: 17134340]

Tate, DC.; Redding, RE. Mental health and rehabilitative services in juvenile justice: System reforms and innovative approaches.. In: Heilbrun, K.; Goldstein, NES.; Redding, RE., editors. Juvenile delinquency: Prevention, assessment, and intervention. Oxford University Press; New York, NY: 2005. p. 134-160.

Vogt DS, King DW, King LA. Focus groups in psychological assessment: Enhancing content validity by consulting members of target population. Psychological Assessment. 1995; 16:231–243. doi: 10.1037/1040-3590.16.3.231. [PubMed: 15456379]

Wright BD. Comparisons require stability. Rasch Measurement Transactions. 1996; 10:506.

Wright, BD.; Stone, MH. Best test design. Mesa Press; Chicago: 1979.

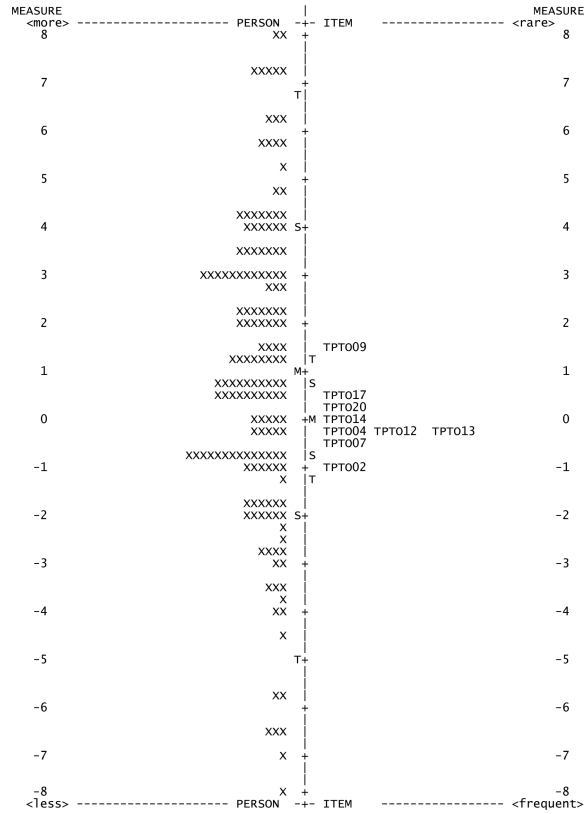

Figure 1.
Person-Item Map for the T4 TPTO:YAB Caregiver dimension.

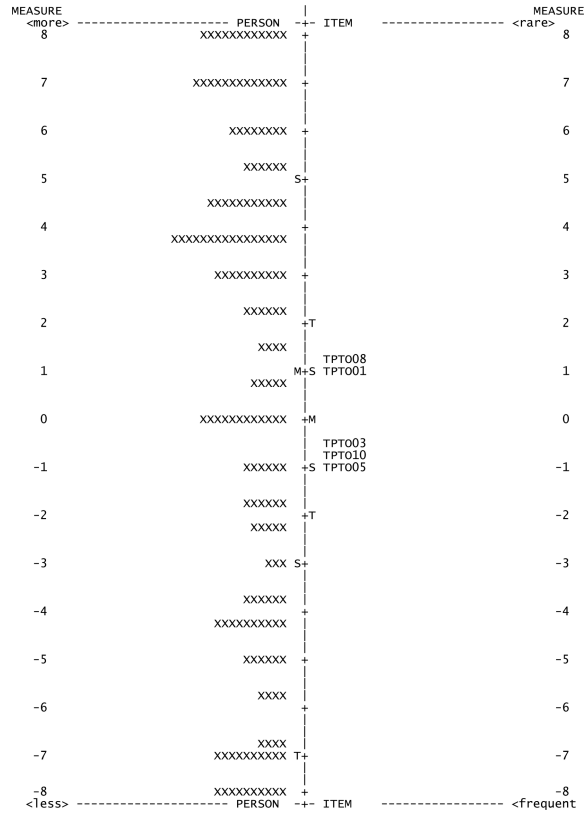

Figure 2.
Person-Item Map for the T4 TPTO:YAB Youth dimension.

Table 1

TPTO:YAB Items

1. The youth continues to engage in the problem behavior(s) that brought him/her into treatment. (Y)
2. The caregivers will do whatever is necessary to help the youth succeed. (C)
3. The family has met the overarching goals of treatment. (Y)
4. The caregivers use what they have learned in therapy in a variety of situations. (C)
5. The youth's problem behavior has improved. (Y)
6. The family needs continued treatment.
7. The caregiver can identify and solve family problems. (C)
8. The youth's behavior places him/her at risk for arrest or placement outside the home. (Y)
9. The caregivers' own life issues keep them from parenting effectively. (C)
10. In general, the youth is doing well. (Y)
11. The caregivers see the problem as belonging to the youth, not to them.
12. The caregivers follow through with what needs to be done to manage the youth. (C)
13. The caregivers believe that they have the skills to improve the youth's behavior. (C)
14. The caregivers can manage their own life issues well enough to parent the youth effectively. (C)
15. The caregivers have given up on the youth.
16. The caregivers communicate well with other systems involved in the youth's life, such as the school.
17. The caregivers consistently use appropriate parenting practices. (C)
18. The youth is doing well in school.
19. The caregivers have a positive attitude about having the youth at home.
20. The caregivers can deal with the youth effectively without needing my (the therapist's) advice. (C)

Note. Items designated "C" were retained for the final Effective Caregiving Outcome scale; "Y" items were retained for the Youth Behavior Outcome scale.

Table 2

Rasch Item Fit Statistics for T4 TPTO:YAB ECO Items

	Infit		Outfit		Item Difficulty		Pt Biseral
Items	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**Measure**	**S.E.**	***r***
2. The caregivers will do whatever is necessary to help the youth succeed.	1.00	.0	1.01	.1	−.99	.15	.82
4. The caregivers use what they have learned in therapy in a variety of situations.	.83	−1.5	.97	−.2	−.25	.15	.83
7. The caregiver can identify and solve family problems.	.70	−2.9	.72	−2.5	−.50	.15	.86
9. The caregivers' own life issues keep them from parenting effectively.	1.73	5.4	1.67	4.2	1.50	.14	.72
12. The caregivers follow through with what needs to be done to manage the youth.	.69	−3.0	.65	−3.2	−.19	.15	.89
13. The caregivers believe that they have the skills to improve the youth's behavior.	1.29	2.3	1.21	1.7	−.28	.15	.76
14. The caregivers can manage their own life issues well enough to parent the youth effectively.	.99	.0	.99	.0	−.12	.15	.83
17. The caregivers consistently use appropriate parenting practices.	.67	−3.2	.73	−2.4	.46	.15	.87
20. The caregivers can deal with the youth effectively without needing my (the therapist's) advice.	.90	−.8	.88	−1.0	.37	.15	.85
Mean	.98	−.4	.98	−.4	.00	.15	
SD	.32	2.7	.29	2.2	.67	.00	

Note. *n* = *162*. MNSQ = mean square (with expectation of 1); ZSTD = standardized mean square fit statistic. Five-point scoring used.

Table 3

Rasch Item Fit Statistics for T4 TPTO:YAB YBO

	Infit		Outfit		Item Difficulty		Pt Biseral
Items	**MNSQ**	**ZSTD**	**MNSQ**	**ZSTD**	**Measure**	**S.E.**	***r***
1. The youth continues to engage in the problem behavior(s) that brought him/her into treatment.	.82	−1.7	.80	−1.8	1.10	.16	.90
3. The family has met the overarching goals of treatment.	1.18	1.5	1.22	1.6	−.49	.17	.88
5. The youth's problem behavior has improved.	.85	−1.3	.82	−1.4	−1.17	.17	.91
8. The youth's behavior places him/her at risk for arrest or placement outside the home.	1.30	2.4	1.20	1.3	1.27	.16	.87
10. In general, the youth is doing well.	.72	−2.5	.75	−2.0	−.72	.18	.92
Mean	.98	−.3	.96	0.5	.00	.17	
SD	.23	1.9	.21	1.6	1.00	.01	

Note. n = 163. MNSQ = mean square (with expectation of 1); ZSTD = standardized mean square fit statistic. Five-point scoring used.

Table 4

Factor Loadings for Confirmatory Bifactor Analyses of TPTO:YAB Items at Times 3 and 4

	TIME 3			TIME 4		
Item	λ_{GEN}	λ_{CG}	λ_{Y}	λ_{GEN}	λ_{CG}	λ_{Y}
1. The youth continues to engage in the problem behavior(s) that brought him/her into treatment.	0.52		0.63	0.80		0.55
2. The caregivers will do whatever is necessary to help the youth succeed.	0.90	0.18		0.80	0.38	
3. The family has met the overarching goals of treatment.	0.69		0.60	0.90		0.31
4. The caregivers use what they have learned in therapy in a variety of situations.	0.87	0.16		0.86	0.3	
5. The youth's problem behavior has improved	0.68		0.57	0.85		0.45
7. The caregiver can identify and solve family problems.	0.88	0.18		0.89	0.31	
8. The youth's behavior places him/her at risk for arrest or placement outside the home.	0.60		0.65	0.73		0.59
9. The caregivers' own life issues keep them from parenting effectively.	0.67	0.58		0.54	0.70	
10. In general, the youth is doing well.	0.57		0.82	0.84		0.48
12. The caregivers follow through with what needs to be done to manage the youth.	0.84	0.36		0.84	0.45	
13. The caregivers believe that they have the skills to improve the youth's behavior.	0.80	0.11		0.83	0.27	
14. The caregivers can manage their own life issues well enough to parent the youth effectively.	0.73	0.67		0.65	0.68	
17. The caregivers consistently use appropriate parenting practices.	0.85	0.37		0.79	0.52	
20. The caregivers can deal with the youth effectively without needing my (the therapist's) advice.	0.81	0.22		0.88	0.31	

Note. λ_{GEN} = factor loading for the general factor, λ_{CG}, λ_{Y} = loading for specific CG and Y factors.

Table 5

Sample Sizes, Means, and Standard Deviations for All Measures

	T3			T4		
	N	M	SD	N	M	SD
TPTO:YAB Total	111	3.30	.78	162	3.31	.95
TPTO:YAB ECO	111	3.28	.83	162	3.34	.91
TPTO:YAB YBO	111	3.32	.93	163	3.26	1.25
APQ Inconsistent Discipline	103	3.56	.71	153	3.58	.64
APQ Poor Monitoring	103	3.68	.70	152	3.59	.69
SRD General Delinquency	100	1.95	2.83	146	2.17	3.76
SIPA Incompetence/Guilt	103	2.47	.79	153	2.50	.78
CBCL Externalizing	103	15.67	12.96	153	15.36	13.43
CBCL Internalizing	103	8.06	7.77	153	7.78	7.94

Note. TPTO:YAB = Therapist Perception of Treatment Outcome: Youth Antisocial Behaviors; ECO = Effective Caregiver Outcomes; YBO = Youth Behavior Outcomes; items scored on 5-point scale; APQ = Alabama Parenting Questionnaire; SRD = Self-Report Delinquency; SIPA = Stress Index for Parents of Adolescents; CBCL = Child Behavior Checklist; T3 = Time 3 (mid-treatment), T4 = Time 4 (at termination). Therapists of families who completed treatment prior to reaching the T3 administration period provided T4 TPTO:YAB data but not T3 TPTO:YAB data.

Table 6

Multilevel Validity Correlations at T3 and T4

	TPTO:YAB-Tot		TPTO:YAB-ECO		TPTO:YAB-YBO	
Measure and Scale	**T3**	**T4**	**T3**	**T4**	**T3**	**T4**
APQ Inconsistent Discipline	.37**	.33**	.24*	.27**	.42**	.22**
APQ Poor Monitoring	.34**	.34**	.17	.25*	.43**	.37**
SRD General Delinquency	−.06	−.27**	.04	−.17*	−.20	−.32**
SIPA Incompetence/Guilt	−.21*	−.19**	−.12	−.22*	−.27**	−.13
CBCL Externalizing	−.20*	−.45**	−.07	−31**	−.33**	−.52**
CBCL Internalizing	−.05	−.42**	−.02	−.31*	−.11	−.45**

Note. n = 111 at T3; n = 163 (TPTO:YAB-YBO) and 162 (TPTO:YAB-Tot, TPTO:YAB-ECO) at T4; Corrleations calculated using MPlus type = complex analyses. APQ = Alabama Parenting Questionnaire (caregiver report); SRD = Self-Report Delinquency (youth report); SIPA = Stress Index for Parents of Adolescents (caregiver report); CBCL = Child Behavior Checklist (caregiver report); T3 = Time 3 (mid-treatment), T4 = Time 4 (post-treatment). Higher scores on APQ reflect better parenting.

* $p < .05$.

** $p < .01$.

Table 7

Zero-Order Validity Correlations at T3 and T4

	TPTO:YAB-Tot		TPTO:YAB-ECO		TPTO:YAB-YBO	
Measure and Scale	**T3**	**T4**	**T3**	**T4**	**T3**	**T4**
APQ Inconsistent Discipline	.29**	.23**	.22*	.22**	.31**	.20*
APQ Poor Monitoring	.21*	.34**	.13	.27**	.27**	.38**
SRD General Delinquency	−.06	−.19*	.03	−.11	−.18	−.25**
SIPA Incompetence/Guilt	−.20*	−.21**	−.13	−.20*	−.24*	−.19*
CBCL Externalizing	−.19	−.29**	−.09	−.18*	−.29**	−.38**
CBCL Internalizing	−.11	−.32**	−.06	−.24**	−.15	−.36**

Note. n = 111 at T3; n = 163 (TPTO:YAB-YBO) and 162 (TPTO:YAB-Tot, TPTO:YAB-ECO) at T4; Correlations calculated using SPSS v. 20. APQ = Alabama Parenting Questionnaire (caregiver report); SRD = Self-Report Delinquency (youth report); SIPA = Stress Index for Parents of Adolescents (caregiver report); CBCL = Child Behavior Checklist (caregiver report); T3 = Time 3 (mid-treatment), T4 = Time 4 (post-treatment). Higher scores on APQ reflect better parenting.

* $p < .05$.

** $p < .01$.