Journal of Substance Abuse Treatment 76 (2017) 1–10

Contents lists available at ScienceDirect

Journal of Substance Abuse Treatment

Regular articles

Predictive validity of an observer-rated adherence protocol for multisystemic therapy with juvenile drug offenders

Marie L. Gillespie [a,*], Stanley J. Huey Jr. [a,b], Phillippe B. Cunningham [c]

[a] Department of Psychology, University of Southern California, 3620 McClintock Avenue, SGM 501, Los Angeles, CA 90089, United States
[b] Department of American Studies and Ethnicity, University of Southern California, 3620 South Vermont Avenue, KAP 462, Los Angeles, CA 90089, United States
[c] Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, 67 President Street Charleston, SC 29425, United States

ARTICLE INFO

Article history:
Received 7 June 2016
Received in revised form 29 December 2016
Accepted 5 January 2017

Keywords:
Multisystemic therapy
Treatment adherence
Independent raters
Substance use

ABSTRACT

Objective: Multisystemic therapy (MST) is perhaps the best validated treatment for youth who engage in serious and chronic antisocial behavior (Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009). Despite evidence suggesting that high treatment adherence is needed to achieve optimal MST outcomes, this research is limited because past studies have relied on adherence reports derived solely from treatment participants (*i.e.*, caregivers, youth, and therapists). To address this gap in the literature, the present study assessed the reliability and predictive validity of an observational protocol for rating adherence to MST.
Method: The sample was drawn from a randomized clinical trial of juvenile drug offenders (77.5% male, 65% African American) referred to one of four treatment conditions (Henggeler et al., 2006). Audiotaped sessions of youth and their families were selected from the first month of MST and trained undergraduate students independently rated therapist adherence to the nine MST treatment principles. We assessed the validity of MST adherence in predicting outcomes at post-recruitment and 12-month follow-up.
Results: Good interrater reliability (ICC = 0.642) was found across all raters for our composite index of adherence. High adherence to MST during the first month of therapy predicted decreases in externalizing behavior at post-recruitment and decreases in youth alcohol consumption at 12-month follow-up.
Conclusions: These results provide independent support for the link between treatment fidelity and behavioral outcomes in the context of MST. Further, this study demonstrates the feasibility of using novice, undergraduate judges to reliably code therapist adherence.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Multisystemic therapy (MST; Henggeler, Schoenwald, Borduin, Rowland, & Cunningham, 2009) is perhaps the best validated treatment for serious and chronic juvenile offending. With nearly 20 published randomized trials supporting its effectiveness (Henggeler, 2011), MST is based on Bronfenbrenner's (1979) model of social-ecology, which argues that child behaviors (*e.g.*, delinquent) develop and are maintained by transactions within and across multiple systems in which the youth is embedded. As such, MST interventions target multiple levels of the youth's social context (*e.g.*, family, peer, school) using strategies that are tailored to the unique needs of each youth. Nine treatment principles guide therapist behavior and selection of specific interventions (Henggeler et al., 2009).

The MST principles (Table 1) require therapists to identify the 'fit' between the youth's presenting problems and their broader systemic context (Principle 1) and to target sequences of behavior within and between multiple systems (Principle 5). Further, therapists use developmentally appropriate (Principle 6), present- focused, action-orientated and well-defined interventions (Principle 4) that are designed to maintain therapeutic change over time (Principle 9). These principles have been said to "serve as the foundation for the MST model and a template against which all interventions can be compared to judge fidelity" (p.13; Henggeler et al., 2009).

Despite strong empirical support, MST outcomes are not uniformly positive, and several studies report poorer effects, particularly for trials conducted in "real-world" treatment settings with novice MST clinicians and less "quality assurance" (*e.g.*, Henggeler, Pickrel, & Brondino, 1999). Some argue that such attenuated treatment effects in effectiveness contexts may result when therapists fail to adhere properly to MST treatment principles. Indeed, several studies indicate that MST effects on behavior problems, drug use, and arrest rates are diminished when therapist adherence varies substantially (Henggeler, Melton, Brondino, Scherer, & Hanley, 1997; Huey, Henggeler, Brondino, & Pickrel, 2000; Schoenwald, Carter, Chapman, & Sheidow, 2008). Thus, treatment fidelity is important to consider when examining the impact of MST on various treatment populations.

* Corresponding author.
E-mail address: marie.gillespie@usc.edu (M.L. Gillespie).

http://dx.doi.org/10.1016/j.jsat.2017.01.001
0740-5472/© 2017 Elsevier Inc. All rights reserved.

Table 1
Multisystemic therapy (MST) treatment principles.

MST principle	Description
Principle 1	Finding the fit: the primary purpose of assessment is to understand the 'fit' between the identified problems and their broader systemic context.
Principle 2	Positive and strength focused: therapeutic contacts should emphasize the positive and should use systemic strengths as levers for change.
Principle 3	Increasing responsibility: interventions are designed to promote responsible behavior and decrease irresponsible behavior among family members.
Principle 4	Present-focused, action-orientated and well-defined: interventions are present-focused and action-orientated, targeting specific and well-defined problems.
Principle 5	Targeting sequences: interventions target sequences of behavior within and between multiple systems that maintain identified problems.
Principle 6	Developmentally appropriate: interventions are developmentally appropriate and fit the developmental needs of the youth.
Principle 7	Continuous effort: interventions are designed to require daily or weekly effort by family members.
Principle 8	Evaluation and accountability: interventions effectiveness is evaluated continuously from multiple perspectives, with providers assuming accountability for overcoming barriers to successful outcomes.
Principle 9	Generalization: interventions are designed to promote treatment generalization and long term maintenance of therapeutic change by empowering caregivers to address family members' needs across multiple systemic contexts.

However, past studies on MST adherence were limited in that they relied exclusively on informants who were directly involved in the treatment process–caregivers, youths, and therapists themselves (Henggeler, 2011). Although including therapy participants in process evaluations is relatively convenient and efficient (Drapeau, 2014), this approach is susceptible to several forms of rater bias (Goense, Boendermaker, van Yperen, Stams, & van Laar, 2014; Perepletchikova & Kazdin, 2005). Youth and caregiver ratings of treatment fidelity may simply reflect their general impressions of the clinician or perceptions of how well they are doing in treatment (Bechger, Maris, & Hsiao, 2010; Kozlowski & Kirsch, 1987). Therapists are also vulnerable to rater bias as they tend to view themselves as more adherent than do independent raters (Breitenstein et al., 2010). Common method variance bias (Campbell & Fiske, 1959; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Reio, 2010) is another potential concern as correlations between MST adherence and treatment outcomes may have been inflated in prior studies because ratings for both constructs were often completed by the same informants. Using independent judges to rate MST treatment adherence should minimize the effects of such biases (Bechger et al., 2010; Goense et al., 2014).

The importance of observational ratings has been considered by other investigators as well, with several studies finding significant associations between observer-rated treatment adherence and reductions in problem behavior for treatment-referred adolescents (Hogue et al., 2008; Robbins et al., 2011). Moreover, one published trial has incorporated observational ratings of MST fidelity, but this was done within the context of health-related outcomes in diabetic youth (Ellis, Naar-King, Templin, Frey, & Cunningham, 2007). However, an observational MST protocol has yet to be validated for delinquent and substance-abusing youth, which is the primary populations targeted by MST.

The present study seeks to address this gap in the MST adherence literature by establishing the reliability and predictive validity of an observational coding protocol using independent raters with no prior knowledge of MST. Participants were drawn from a randomized trial of juvenile drug offenders who received MST or alternate treatments in the context of juvenile drug court or family court. We hypothesized that higher therapist adherence to the MST treatment protocol, as measured by independent raters, would predict better treatment outcomes.

2. Methods

2.1. Participants

The sample was drawn from a randomized clinical trial assessing the efficacy of MST and drug court for juvenile drug offenders (Henggeler et al., 2006). Families were recruited from the Department of Juvenile Justice (DJJ) in Charleston County, South Carolina, and all youth met DSM-IV diagnostic criteria for alcohol or drug abuse or dependence (APA, 1994). The initial trial included 161 juvenile drug offenders who were randomly assigned to one of four treatment conditions: Family court (FC), drug court (DC), drug court plus MST (DC/MST), or drug court plus MST with contingency management (DC/MST/CM). The current study included 40 youth and their caregivers from the DC/MST and DC/MST/CM conditions whose therapy sessions were audio-recorded. The mean age for youth in this subsample was 15.4 years ($SD = 1.1$), and most were male (77.5%) and African American (65.0%) or Caucasian (32.5%). The median family income was in the $10,000–$15,000 range and over a third of families were receiving financial assistance. All procedures for Human Subjects research were approved by the institutional review board at the Medical University of South Carolina.

2.2. Therapists and intervention

MST was provided by six master's-level therapists with degrees in social work, psychology, or education. All were female, three were African American, and three were European American. Consistent with MST's model of service delivery, therapists provided in-home services, were available 24 h a day, 7 days a week, and interacted with families from 2 to 15 h a week as needed. The MST quality-assurance framework (Schoenwald, Sheidow, & Letourneau, 2004) was also implemented; this included site assessments, ongoing therapist training and consultation, review of audio-recorded sessions in weekly supervision of MST teams, and qualitative and quantitative assessment of therapist adherence. The training of clinicians, supervisors, and organizations were manualized and quarterly booster sessions were provided to monitor for adequate clinical competencies. Working closely with caregivers, therapists emphasized family strengths, and provided skills to cope with disruptive or maladaptive behaviors. In addition to the services provided in DC/MST, the DC/MST/CM condition incorporated components of contingency management. More information on procedures in the randomized trial is available elsewhere (Henggeler et al., 2006).

2.3. Procedures

2.3.1. Coder training

Coders were three undergraduate students with no prior exposure to MST. Over a period of nine weeks, readings (Henggeler et al., 2009), graded learning tasks, and sample audiotapes of in-home therapy sessions were used to familiarize coders with MST and the coding protocol (Huey, 2001). Our coders rated seven training tapes using their knowledge of the protocol, and ratings were compared to "gold standard" ratings provided by the first and second authors. Discrepant ratings were discussed in group meetings that were held twice a week to monitor coder drift. At the end of the training process, coders achieved acceptable reliability (ICCs = 0.830) and consistency (64–79% agreement).

2.3.2. MST coding

Using a random number generator, coders were organized into pairs to independently rate the 40 selected sessions. To monitor and address rater drift, weekly meetings were held and discrepant ratings for every

session were discussed. Moreover, every third week, all three coders were given the same session to rate in addition to their regular assignments; their ratings were compared to the expert rating and discrepancies were discussed.

Sessions were randomly selected from the first month of treatment, as early sessions appear to be more predictive of treatment outcomes and more representative of the therapy process (Gomez-Schwartz, 1978; Suh, Strupp, & O'Malley, 1986). Of the audio-recorded sessions available for each family, at least one session had to meet all of the following criteria: (1) recorded within the first month of treatment, (2) more than 15 min and less than 90 min in length, (3) reasonably audible (*i.e.*, no severe static, background noise, or muffled voices), and (4) involved therapy for most of the session (*e.g.*, sessions wherein the family mostly completed surveys were excluded). Of the 81 MST families, only 40 met all of these conditions. These sessions were then randomly assigned to coder dyads and coded in their entirety. When more than one recorded session met our four criteria for a particular family, one session was randomly selected and assigned. Session length ranged from 23 min to 78 min ($M = 44$ min), with sessions recorded between 5 days and 31 days ($M = 16$ days) after the start of treatment. Depending on the needs of a particular family, therapy attendees varied from session to session and included a combination of caregivers, siblings, and target youth. Therapists were evaluated on how well they adhered to MST principles when interacting with these various family members during sessions.

2.4. Measures

2.4.1. MST adherence

Coders listened to audio-recorded therapy sessions in their entirety and evaluated therapist adherence to the nine MST principles using a 30-item coding protocol rated on a 6-point scale. The coding manual included specific descriptors and examples that represent low (1–2), moderate (3–4), and high (5–6) therapist adherence to the MST principles. Each principle was represented by two to four "subprinciple" items and one "overall" item; total session adherence was calculated by taking the average of the nine "overall" items, thus creating an "MST Principle Adherence" (MPA) composite.

For example, Principle 1 (*The primary purpose of assessment is to understand the "fit" between the family-identified problems and their broader systemic context*) was composed of Subprinciple 1 *(The therapist understood what key factors contributed to the target problems(s) or issues; i.e., understood "fit"*), Subprinciple 2 (*The therapist tapped all sources needed to appropriately evaluate "fit"*), and an Overall Principle score (*Overall, this principle was properly implemented*) that reflected the conceptual average of the two subprinciples. The MPA composite therefore represents the mean of all MST principles.

To illustrate how therapist adherence to each individual MST principle was assessed within treatment sessions, in Appendix A we present descriptions of therapist-client interactions that reflect desired behavioral indicators for each coding category and level. These excerpts were taken from the coding manual provided to trained raters. Examples illustrate how well a therapist addresses the target child's (TC) core problems throughout the entirety of the session.

For example, regarding Principle 1, Subprinciple 1, a therapist might receive a low rating if he or she learned that TC violated curfew but made no effort to obtain further information from the parents regarding the antecedents or consequences of that behavior. The therapist might receive a moderate rating if he or she evaluated factors that anteceded TC's problem behavior (*e.g.*, after inquiries, it is determined that TC had a fight with mom prior to leaving the house), but did not attempt to evaluate the consequences of that behavior. A high rating might be given if the therapist delved into the specific antecedents (*e.g.*, TC's aggressive behavior often happens after an argument with his mom) and consequences (*e.g.*, mom does not enforce the agreed upon consequence of TC losing television privileges) surrounding TC's problem behavior.

2.4.2. Outcome measures

Youth and caregiver ratings were collected at pretreatment (T1; conducted within 3 days of the youth's entry into the study), post-recruitment (T2; 4 months post-entry), and follow-up (T3; 12 months post-entry). Arrest and urine drug screen data were collected over the course of treatment and aggregated at the indicated time points.

The Child Behavior Checklist (CBCL; Achenbach, 1991) was completed by adolescents and caregivers. This 113-item questionnaire has demonstrated good discriminant validity between referred and non-referred children of various ethnic and age groups. The CBCL evaluates youth externalizing and internalizing problems. *T* scores for the broadband externalizing scale, which yielded Cronbach's alpha reliabilities of 0.92 in previous studies (Gross et al., 2006), were used for the current study.

Youth also completed the Self-Report Delinquency Scale (SRDS; Elliott, Ageton, Huizinga, Knowles, & Canter, 1983), which is one of the best validated delinquency scales and has exhibited good test–retest reliability (Thornberry & Krohn, 2000). This 40-item questionnaire examines a range of behaviors from minor (*e.g.*, cheated on school tests) to major (*e.g.*, had/tried to have sexual relations with someone against their will) acts of delinquency committed during the previous three months. The General Delinquency subscale was used for outcome analyses.

Youth arrests were tracked through computerized records housed at the South Carolina DJJ. For youth over the age of 16 years, adult criminal records were also collected from the South Carolina Law Enforcement Division.

Youth completed the Form 90 (Miller, 1996), an interview based on the time line follow-back (TLFB) methodology to estimate specific amounts of alcohol and other drugs consumed on a daily basis. The TLFB approach has been found to be a reliable assessment of adolescent substance use, as self-reports correlated with biological markers and collateral reports in previous validation studies (Donohue et al., 2004; Waldron, Slesnick, Brody, Turner, & Peterson, 2001). Total days of alcohol use and total days of marijuana use were used for outcome analyses.

Urine drug screens for cannabis were collected using the 3-Test Integrated Cup supplied by BioTechNostix (Markham, Ohio) before each drug court appearance. Drug screens were dichotomized as positive (one or more positive drug screens) or negative (no failed drug screens) at post-recruitment and 12 months follow-up. In line with juvenile drug court protocols, youths with unexcused absences (*e.g.*, did not show, runaway) and youths recently placed in detention, thereby missing court, were counted as having positive urine screens. Further, juveniles with excused absences (*e.g.*, attending class) were counted as having negative/clean drug screens.

2.5. Analyses

Intraclass correlation coefficients (ICCs) were calculated for all three pairs of coders to estimate reliability. The convention developed by Cicchetti (1994) for evaluating the usefulness of ICCs was adopted: below 0.40 = poor, 0.40 to 0.59 = fair, 0.60 to 0.74 = good, and 0.75 to 1.00 = excellent.

Because several outcomes represented a type of count variable (*e.g.*, number of drinking days), predictive validity was tested using negative binomial generalized estimating equations (GEE; Liang & Zeger, 1986). GEE deviates from the classic assumption of statistical independence in traditional regression models and calculates the data by estimating a working correlation matrix; this was done in the course of running negative binomial generalized estimating equations. Two-way interactions between the MPA composite and time on the outcome variables were separately evaluated in the presence of baseline covariates. To interpret negative binomial results, coefficients were exponentiated (*i.e.*, $\mathrm{Exp}(\beta)$) and yielded rate ratios (RRs). Similar to odds ratios in logistic regression, an RR value larger than 1 indicates a percentage increase

in counts for each unit increase in the predictor, and a value less than 1 indicates a percentage decrease in the outcome for each unit decrease in the predictor. Additionally, a GEE linear model was used for the normally distributed CBCL data, and logistic regression analyses were used for dichotomized data (*e.g.*, arrest data).

Youth and primary caregiver sex (male) and race (African American) were associated with attrition patterns in outcome data and were added as covariates in our models. The relationship between MPA to each outcome did not differ significantly between the adjusted and non-adjusted models. Thus, the following results represent data from the non-adjusted models.

3. Results

3.1. Descriptive statistics

The sample sizes, means, standard deviations, and ranges for all outcome variables are presented in Table 2. Most outcome variables were positively skewed, indicating that less severe problem behavior was more common than severe problem behavior. For analytic purposes, both youth arrest and drug screen data were dichotomized as either 0 (no arrest, all negative drug screens) or 1 (one or more arrests, one or more positive drug screens). This method is commonly used in prevention research when variables are not truly continuous and skewed with infrequent data greater than 0 (Farrington & Loeber, 2000). Similar to the original sample (Henggeler et al., 2006), means for most outcomes in this study decreased from pre-treatment to follow-up.

With regard to treatment fidelity, the six therapists in this trial showed similar levels of high adherence, with MPA mean levels ranging from 4.54 to 4.77 (possible MPA range = 1.00–6.00) across sessions. The distribution of MPA levels across sessions is presented in Table 3. No significant group differences in adherence were found between the DC/MST ($n = 14$) and DC/MST/CM ($n = 26$) conditions, $t(38) = -0.10$; $p = 0.92$, with both conditions combined yielding average MPA ratings of 4.6 ($SD = 0.34$). Further, Chi Square analyses showed no differences in demographic variables or youth behavior severity (*e.g.*, arrest) between families that were selected ($n = 40$) and not selected ($n = 41$) for the current study. However, DC/MST/CM families were significantly more likely than DC/MST families to be selected for inclusion in the final sample, $\chi^2(1) = 4.50$, $p = 0.03$.

3.2. Predictive validity

Reliability of the adherence ratings for each of the 40 families varied from fair (ICC = 0.549) to excellent (ICC = 0.759) between the rater dyads for MPA, and was good when calculated across all raters (ICC = 0.642). Therefore, MPA was used as the predictor of youth outcomes. A negative binomial GEE with unstructured covariance pattern was fit to each outcome for count data. Predictors included time (T1, T2, and T3) and MPA. GEE results are presented in Table 4 and logistic regression results are presented in Table 5. A significant effect of MPA on changes in alcohol use was found ($\text{Exp}(\beta) = 0.022$, 95% CI [0.002, 0.214]); specifically, number of days of alcohol use decreased by 98% from T1 to T3 for youth involved in sessions that were rated as highly adherent to MST (*i.e.*, MPA over 4.5).

Further, a GEE linear model with unstructured covariance pattern was fit to CBCL outcome data. Predictors included time and MPA. There was a significant effect of MPA on youth-reported externalizing T-scores from T1 to T2 ($B = -0.8.804$, 95% CI [−15.532, −2.077]); sessions with high MPA ratings (*i.e.*, over 4.5) predicted a nearly 9-point decrease in youth-reported externalizing symptoms at T2. No significant effects were found for arrest, delinquency, marijuana use, or parent ratings on the CBCL.

Table 3
Distribution of ranges and sample sizes for MPA.

MPA range	N	%
3.5–4.0	2	5
4.1–4.5	10	25
4.6+	28	70

Note: MPA = MST Principle Adherence.

4. Discussion

Despite the extensive use of client ratings to assess MST fidelity in past studies (Henggeler & Borduin, 1992; Henggeler et al., 1997; Schoenwald et al., 2008), adherence to MST has yet to be systematically assessed using independent raters. Recent efforts have been made to address this gap in the adherence literature. In the context of a randomized trial of MST, Weiss et al. (2013) had a single independent expert in MST review audiotaped sessions and rate adherence to the nine principles, with the clinical goal of assuring adequate fidelity during treatment. While the authors acknowledged the importance of the principles, they did not report on the reliability or validity of those ratings.

Our coding system (Huey, 2001) was used in a previous study of MST with diabetic youth (Ellis et al., 2007) and observational ratings of adherence were found to correlate with health-related outcomes. The protocol was revised and shortened by the first author in order to improve rater reliability. The current study is the first to assess the reliability and predictive validity of an observational protocol for assessing adherence to MST with juvenile offenders, who are the primary recipients of MST (Henggeler, 2011; Henggeler et al., 2009). Undergraduate coders reliably rated adherence to MST during the first month of treatment, and higher adherence was predictive of significant decreases in

Table 2
Means, standard deviations, and sample sizes for outcome variables.

Outcome variable	T1[a]			T2[a]				T3[a]			
	Mean	*SD*	*n*	Mean	*SD*	*n*	% miss	Mean	*SD*	*n*	% miss
CBCL externalizing T-score											
Caregiver report	62.75	10.52	40	55.36	12.44	39	2.5	51.94	13.27	35	12.5
Youth report	59.13	12.87	40	52.62	14.79	37	7.5	50.69	13.51	36	10
Form 90 – Alcohol	5.18	12.01	40	1.05	4.62	40	0	1.25	6.50	36	10
Form 90 – Marijuana	28.30	27.27	40	2.75	7.97	40	0	1.67	4.84	36	10
SRDS (General Delinquency)	30.95	37.78	40	20.74	30.95	38	5	15.67	28.19	36	10
Outcome variable					%	*n*	% miss		%	*n*	% miss
Arrests[b,c]					35	14	0		50	20	0
Urine screens – cannabis[c]					65	24	7.5		51	18	12.5

Note: CBCL = Child Behavior Checklist; SRDS = Self-Report Delinquency Scale; % miss = percentage of missing data.
[a] T1 = pretreatment, T2 = 4-month post-recruitment, T3 = 12-month post-recruitment.
[b] Arrests represent all types of arrest (drug and violent).
[c] Arrest and Drug data represent dichotomous proportions with "%" representing percentage and "*n*" representing the number of youth in the sample with one or more arrest or one or more positive drug screen.

Table 4
Generalized estimation equation (GEE) model estimates of the relationship between MST Principle Adherence and outcome variables across time.

Outcome variable	MPA * T1–T2[a]				MPA * T1–T3[a]			
		95% CI				95% CI		
	B	Lower	Upper	*p*-Value	B	Lower	Upper	*p*-Value
Scale – Linear								
CBCL externalizing T-score[a]								
Caregiver report	−0.383	−8.444	7.677	0.926	4.954	−3.143	13.050	0.230
Youth report	−8.804	−15.532	−2.077	0.010	−4.017	−12.220	4.186	0.337
Counts-Negative Binomial		95% CI				95% CI		
	Exp(β)	Lower	Upper	*p*-Value	Exp(β)	Lower	Upper	*p*-Value
Form 90 – Alcohol	0.405	0.011	14.932	0.624	0.022	0.002	0.214	0.001
Form 90 – Marijuana	0.352	0.021	5.941	0.469	7.614	0.678	85.506	0.100
SRDS (General Delinquency)	0.731	0.207	2.587	0.627	0.437	0.139	1.375	0.157

Note: CBCL = Child Behavior Checklist; SRDS = Self-Report Delinquency Scale; * = interaction.
[a] T1 = pretreatment, T2 = 4-month post-recruitment, T3 = 12-month post-recruitment.

several, although not all, domains of youth problem behavior. Specifically, youth who were in high-adherence sessions (*i.e.*, MPA of over 4.5) early in treatment decreased their alcohol consumption by 98% from pre-treatment to 12-month follow-up. MPA was also associated with reductions in youth-rated externalizing symptoms. Youth involved in high-adherence sessions reported a nearly 9-point decrease in externalizing symptoms on the CBCL from pre-treatment to post-entry. Although an adherence association was not observed for *caregiver* CBCL ratings, poor cross-informant correlations (*i.e.*, youth *versus* caregiver reports of symptoms) are not uncommon in the child intervention literature for this measure (see Achenbach, McConaughy, & Howell, 1987; Rescorla et al., 2013).

These results provide additional support for the link between fidelity and select treatment outcomes in the context of MST (Chapman & Schoenwald, 2011; Henggeler et al., 1997; Huey et al., 2000; Schoenwald, Henggeler, Brondino, & Rowland, 2000; Schoenwald et al., 2008). Prior MST studies were limited in that they relied solely on informants who were directly involved in the therapy process, which presents various biases when assessing adherence (Bechger et al., 2010; Breitenstein et al., 2010; Horenstein, Houston, & Holmes, 1973; Kozlowski & Kirsch, 1987; Podsakoff et al., 2003; Reio, 2010). An important contribution of the current study is our finding that MST adherence as rated by independent judges is predictive of key treatment outcomes for youth with multiple behavioral problems. Moreover, because coders were undergraduates with no clinical experience and minimal prior exposure to MST, our results suggest that novices are capable of evaluating adherence to complex psychosocial interventions such as MST (Baker, Haltigan, Brewster, Jaccard, & Messinger, 2010; Waldinger, Schulz, Hauser, Allen, & Crowell, 2004).

However, future iterations of this protocol should focus on improving the coding training process by addressing disagreements on more complex MST principles. Principles 1 and 4 were the most consistently reliable indices across dyads, but Principles 5 and 9 yielded the poorest reliability (Appendix B). During weekly coding meetings, opinions differed regarding the amount of therapist effort needed to provide community resources for a family in the early stages of treatment (*i.e.*, Principle 9: long-term maintenance of therapeutic change). Further, two of our undergraduate coders appraised therapist-caregiver interactions very differently with regard to the level of attention provided in response to a treatment barrier (*e.g.*, ignoring *versus* addressing a difficult youth's stubbornness) in relation to Principle 5 (*i.e.*, interventions targeting sequences of behavior). Consequently, these types of divergent ratings lowered the reliability of a few principles and further emphasize the need for monitoring coder drift beyond the training phase. For this reason, in part, the MPA composite was found to be the most useful index of adherence for the current study.

Several limitations should be noted. First, although our significant effects were in the hypothesized direction, causality from adherence to outcomes can only be inferred. For example, given that youth in high-adherence sessions drank considerably more than youth in low-adherence sessions at pre-treatment and showed sharper decreases in drinking at follow-up, directionality is unclear. It may be that therapists were responding to extreme levels of initial drinking with higher rates of MST fidelity. Conversely, heavy drinkers may be more responsive to high-adherence therapists compared to youth who drink less. While no research currently exists on the direct association between baseline *drinking rates* and subsequent MST adherence, caregiver-rated fidelity may be lower when youth display more severe pre-treatment *antisocial behavior* (Schoenwald, Halliday-Boykins, & Henggeler, 2003).

Second, in addition to the restrictions our small sample size presents, another limitation was a "ceiling effect" observed in the homogeneity of MST skill in our clinician sample, with adherence rated relatively high across therapists and sessions. While high fidelity is ideal when providing services to high-needs families, skill diversity across therapists is preferable when attempting to establish the reliability and validity of therapy process and observational coding protocols (Breitenstein et al., 2010). For this reason, our coding system might be most useful when used with novice MST therapists in "real-world" treatment

Table 5
Logistic regression of dichotomized data with MST Principle Adherence across time.

Outcome variable	T1–T2[a]				T2–T3[a]			
	95% CI				95% CI			
	Exp(β)	Lower	Upper	*p*-Value	Exp(β)	Lower	Upper	*p*-Value
Arrest[b]								
0 arrest/1 + arrest	1.124	0.157	8.036	0.908	9.938	0.242	908.003	0.226
0–1 arrest/2 + arrests	15.517	0.103	2338.847	0.284	0.754	0.092	6.179	0.792
Drug screens (positive/negative)								
Cannabis	3.754	0.494	28.519	0.201	1.400	0.176	11.152	0.751

Note: Drug screens dichotomized as negative/clean drug screens *vs.* one or more positive/dirty drug screens.
[a] T1 = pretreatment, T2 = 4-month post-recruitment, T3 = 12-month post-recruitment.
[b] Arrests represent all types of arrest (drug and violent).

settings; as such, clinicians may show the greatest variation in skills but the greatest potential for improvement. Although self-reported ratings of adherence are arguably conceptually different than observational ratings, it should be noted that previous MST studies have also found consistently "high" therapist adherence to the MST protocol as reported by informants directly involved in treatment (*e.g.*, Letourneau, Sheidow, & Schoenwald, 2002).

Our current limitations present interesting avenues for future research. Although it goes beyond the scope of the current paper, it would be useful to assess the predictive validity of adherence to each individual principle and to use sessions that span the course of MST treatment. Lastly, the contingency management (CM) component in the DC/MST/CM condition warrants further examination. In the original trial, Henggeler et al. (2006) reported the marginally stronger effect CM had in reducing youth substance use compared to the "pure" MST condition. With a larger sample size, a CM by adherence interaction may be observed with the current study's observational protocol.

5. Clinical implications and conclusion

The MST quality-assurance process aims to implement evidence-based interventions with the fidelity needed to produce positive treatment outcomes for youth and their families (Schoenwald et al., 2004). Our revised coding system, once replicated with a larger and more diverse study sample, has the potential to be used within this quality-assurance framework. Providing clinicians with detailed feedback, such as guiding them to be more mindful of a specific MST principle for which they had lower adherence ratings, could be beneficial in strengthening the implementation of MST. As shown in the current study, the difference between an MPA of 4.0 and 4.5 can result in very different impacts on youth alcohol consumption and behavioral outcomes. It may therefore be important to train therapists to reach a certain threshold level of MST adherence and to monitor fidelity over time in order to ensure optimal outcomes in non-research settings.

In conclusion, this is the first study to evaluate the predictive validity of an observational measure of MST adherence for juvenile offenders. Our findings further illustrate the benefits of well-implemented MST for substance-abusing youth. While demonstrating the feasibility of using novice, undergraduate judges to reliably code therapist adherence, the current study provides independent support for the link between MST fidelity and youth outcomes.

Acknowledgments

The research described in this article was supported by Grants R01AA012202 from the National Institute on Alcohol Abuse and Alcoholism, H79TI14150 from the Substance Abuse and Mental Health Services Administration/Center for Substance Abuse Treatment, and DA013066 from the National Institute on Drug Abuse awarded to Scott W. Henggeler, Ph.D. (Principal Investigator). We sincerely thank Nicholas Jackson for his statistical consultation, and Joseph Green, Sebastienne Leo, and Tasia Mamiya for their assistance in audiotape coding.

Appendix A. Abbreviated coding manual descriptions and examples of MST principles and therapist (Th) rating levels.

P1S1. Therapist understood what key factors contributed to the target problems(s) or issue(s)	
Adherence rating	Description and example(s)
1 or 2 Low	➢ Th has a poor understanding of, or conducts a poor evaluation of, the factors which contribute to TC/family problems, and makes no effort to gather appropriate data for "fit" • *e.g., Th learns that TC violated curfew yesterday, but obtains no data regarding antecedents or consequences; consequence for TC staying out late is simply to lock TC out of home, but TC responds by staying with deviant peers for several days at a time, and no discussion by Th of "fit" or barriers, and no adjustment made to intervention*
	➢ Th offers highly abstract or very distal factors as primary determinants of problem behavior, with no suggestion of more proximal mediating links • *e.g., Th notes how TC's oppositional behavior results from anger he harbors toward mother for leaving his father 5 years ago*
3 or 4 Moderate	➢ Th makes a significant but inadequate/incomplete/partial evaluation of "fit" • *e.g., antecedents of TC's recent drug use evaluated, but not relevant consequences* ➢ Th understands "general" but not specific contributors to TC/-family problems • *e.g., Th explains that peers can be an important factor in children's oppositional behavior, but does not discuss or evaluate factors specific to TC's fights at school*
5 or 6 High	➢ Th understands the factors which contribute to key TC/family problems, treatment barriers, or treatment success • *e.g., Th summarizes how TC hanging out with delinquent peers and parental permissiveness appear to act as contributors to TC's drinking; Th notes how reduction in TC's drinking was preceded by caregiver rule enforcement* ➢ Th arranges for collection of appropriate data to evaluate hypothesis regarding "fit" • *e.g., caregiver says that TC is cutting school because he does not like his teachers, and Th recommends that caregiver talk with TC and teachers*
P1S2. Therapist tapped all sources needed to appropriately evaluate "fit"?	
1 or 2 Low	➢ Th is lacking information from virtually all sources needed to obtain a good "fit", and indicates no plan to obtain information from these sources • *e.g., caregiver briefly notes that TC was suspended from school, but Th does not discuss antecedents with TC, nor initiate plan to contact school personnel*
3 or 4 Moderate	➢ Th obtains or makes efforts to obtain information from some important sources relevant to "fit" of key target issues, but not other crucial sources • *e.g., TC tests positive on urine screen and discusses with caregiver, but no effort to gather "fit" data from TC about drug use*
5 or 6 High	➢ Th obtains, or has made efforts to obtain, information from all or nearly all sources needed to obtain a good "fit" of key target issues • *e.g., Th obtains data on TC's truancy from TC, caregiver, and stepfather, and makes appointment to obtain perspectives of school personnel*
P2S1. Therapist highlighted, identified, or elicited systemic strengths.	
1 or 2 Low	➢ Th misses opportunities to identify family competencies, efforts, successes, and resources • *e.g., TC states that he has avoided marijuana use, but Th seems to ignore or minimize* ➢ Th "lecturing," showing frustration, inappropriately critical, or using pejorative language • *e.g., Th joins mom in "lecturing" TC on school truancy and says he is a "slacker"*
3 or 4 Moderate	➢ Th identifies vague/ambiguous competencies/efforts, or strengths in a disengaged manner • *e.g., at the end of the session, Th vaguely says "well done today" to TC*
5 or 6 High	➢ Th identifies family competencies, efforts, successes, and resources • *e.g., Th praises TC for honestly admitting to drug use and explains why this is important* ➢ Th uses reframing strategies for a more constructive way to approach problems • *e.g., mother blames herself for TC's behavior, but Th points out the multiple contributors while indicating how mother's involvement is critical to solving TC problems*
P2S2. Therapist made recommendations, or designed interventions that used systemic strengths to facilitate change or maintain an existing intervention.	
1 or 2 Low	➢ Th fails to design, implement, or reinforce interventions that are based on family ideas, skills, competencies, efforts, or successes, or indigenous resources

Appendix A. *(continued)*

P1S1. Therapist understood what key factors contributed to the target problems(s) or issue(s)

- *e.g., depressed caregiver notes how she was less depressed when she had regular conversations with a friend, but rather than use as opportunity to encourage network support building, Th ignores*

3 or 4 Moderate
- ➢ Th designs, implements, or reinforces interventions that are based on family ideas, skills, competencies, efforts, or successes, or indigenous resources, but vague or incomplete
- *e.g., Th states "based on what you've accomplished so far, you can do anything you put your mind to," but offers no further elaboration*

5 or 6 High
- ➢ Th designs, implements, or reinforces interventions that are based on family ideas, skills, competencies, efforts, or successes, or indigenous resources
- *e.g., with Th's encouragement during session, aunt agrees to monitor and reinforce caregiver's drug abstinence*

P3S1. The intervention supported responsible behavior.

1 or 2 Low
- ➢ Th fails to discuss, implement, or reinforce strategies that facilitate behaviors in the youth or caregiver which promote positive outcomes
- *e.g., TC states that he completed his homework for the week, but Th does not reinforce or otherwise support this behavior*

3 or 4 Moderate
- ➢ Th discusses, implements, or reinforces strategies that may facilitate behaviors in the youth or caregiver which promote positive outcomes, but vague, unclear, or underdeveloped
- *e.g., Th suggests that TC join a sports team, but does not provide details on how to do so*

5 or 6 High
- ➢ Th discusses, implements, or reinforces strategies that clearly facilitate behaviors in the youth or caregiver which promote positive outcomes
- *e.g., Th helps to arrange for TC to join neighborhood basketball league which includes predominantly prosocial youth*

P3S2. The intervention discouraged irresponsible behavior.

1 or 2 Low
- ➢ Th fails to challenge, recommend the use of consequences for, or offer alternatives to inappropriate behavior
- *e.g., TC has "dirty" drug screen, but no discussion or implementation of consequences*

3 or 4 Moderate
- ➢ Th challenges, recommends consequences for, or offers alternatives to inappropriate behavior, but unclear, vague, or distal
- *e.g., Th points out how TC's drug use may lead to future arrests, but no apparent plan to address immediate consequences*

5 or 6 High
- ➢ Th appropriately challenges, recommends consequences for, or offers alternatives to inappropriate behavior
- *e.g., Th helps father implement weekend grounding for TC's missed curfew*

P4S1. The intervention/assessment focused on present problems, concerns, or issues.

1 or 2 Low
- ➢ Session addresses past problems or issues, with no apparent relevance to current circumstances
- *e.g., part of session focuses on caregiver's childhood relationship with own parents, with no clear link to current target problems*

3 or 4 Moderate
- ➢ Past problems or issues are addressed in session, but with vague or unclear relevance to current circumstances
- *e.g., Th addresses how caregiver coped with an emotionally distant mother and how this might have influenced her current parenting, but there is no discussion of caregiver's ongoing child-rearing practices*

5 or 6 High
- ➢ Session addresses present or recurrent problems, successes, strategies, or issues, or anticipates future problems
- *e.g., father begins to complain about how TC burned down storage shed last year, but Th quickly redirects to focus on TC's current antisocial behaviors*

P4S2. The problems, issues, or concerns were specific and well-defined.

1 or 2 Low
- ➢ Target problems and treatment goals very unclear, unmeasurable, or not addressed
- *e.g., Th warns caregiver of need to "be firm" with TC, but no further discussion of what specific caregiver behavior is problematic and what steps should be taken to resolve*

3 or 4 Moderate
- ➢ Target problems and or treatment goals are stated, but vague and never clarified by Th
- *e.g., parents say that they feel a lack of support from police, but Th elicits no clarification regarding how lack of support is manifested*

5 or 6 High
- ➢ Target behaviors are clear, concrete, and measurable
- *e.g., Th showed family TC's drug screens plotted over time*

Appendix A. *(continued)*

P1S1. Therapist understood what key factors contributed to the target problems(s) or issue(s)

P4S3. The intervention/assessment required action (*e.g., role-play, charting of problems, exposure, contingency contracting*) during the session.

1 or 2 Low
- ➢ Th is passive or permits family to remain passive during session
- *e.g., mom says she doesn't know how to deal with TC's drinking anymore, but Th simply empathizes without assessing or problem-solving*

3 or 4 Moderate
- ➢ Vague or inadequate action required of family members during the session
- *e.g., Th hands breathalyzer kit to caregiver to use with TC, but provides very limited directions and no demonstration*
- ➢ Th focuses almost solely on relaying information to or assigning homework to family members with minimal effort to practice skills, evaluate comprehension, or elicit feedback
- *e.g., Th informs family of possible side-effects of medication, but does not encourage caregiver/youth to monitor medication effects*

5 or 6 High
- ➢ Th helps family take concrete action during session
- *e.g., Th reviews, discusses, and practices with TC how to respond during drug court*
- ➢ Th and family review ongoing, successful strategies utilized by family members
- *e.g., Th praises mom for success in implementing "rules and consequences" with TC at home and reducing TC's school truancy, and encourages her to continue with progress*

P4S4. The intervention/assessment targeted the defined problems, issues, or concerns.

1 or 2 Low
- ➢ Intervention/assessment not clearly relevant to treatment goals, or with target issues as noted in session
- *e.g., session focuses on father's weight problem with no clear relevance to existing goals*

3 or 4 Moderate
- ➢ Intervention/assessment targets intermediary or overarching goals or target issues as noted in session, but incomplete or vague
- *e.g., Th/family discuss plan to implement unspecified "consequences" to TC if she continues to miss curfew*

5 or 6 High
- ➢ Intervention/assessment requires specific action that addresses target issues as noted in session and/or intermediary/overarching treatment goals
- *e.g., based on treatment goals noted by mom, Th assigns caregiver to record the number of daily noncompliant behaviors exhibited by TC*

P5S1. The intervention/assessment targeted sequences of behavior within systems. (*e.g.*, caregiver-youth, family-family, youth-youth)

1 or 2 Low
- ➢ Intervention/assessment does not address or poorly addresses within-system sequences or contingencies affecting current circumstances/behavior
- *e.g., TC explains that he gets tempted to smoke weed all the time, but Th does not discuss triggers and sequences, or otherwise assess*

3 or 4 Moderate
- ➢ Intervention/assessment addresses or targets within-system sequences or contingencies affecting current circumstances/behavior, but vague or incomplete
- *e.g., Th discusses with caregiver general connection between positive reinforcement and increased prosocial behavior but with no specific application to target issues in family*

5 or 6 High
- ➢ Intervention/assessment addresses or targets clear or, specific within-system sequences or contingencies affecting current circumstances/behavior
- *e.g., Th helps TC link antecedent thoughts and feelings to aggressive behavior*

P5S2. The intervention/assessment targeted sequences of behavior between systems. (*e.g.*, caregiver-school; youth-peer)

1 or 2 Low
- ➢ Intervention or assessment fails to address or target between-system sequences or contingencies affecting behavior
- *e.g., Th conceptualizes aggressive behavior solely in terms of TC's cognitive distortions when evidence of delinquent peer affiliation exists*

3 or 4 Moderate
- ➢ Intervention or assessment addresses between-system contingencies or sequences but vague or incomplete
- *e.g., Th encourages TC to "walk away" from peer provocations at school, but provides no specific strategies*

5 or 6 High
- ➢ Intervention or assessment addresses clear or specific between-system contingencies or sequences affecting current behavior

(continued on next page)

Appendix A. *(continued)*

P1S1. Therapist understood what key factors contributed to the target problems(s) or issue(s)

- *e.g., Th assists caregiver in identifying specific conditions under which TC seems to perform well in school*

P6S1. The intervention's developmental level matched the developmental level of the youth/family.

1 or 2 Low
- ➢ Intervention/assessment involves tasks that are clearly inappropriate, or above or below the physical, emotional, or cognitive capacities and needs of family members
- *e.g., Th encourages developmentally delayed 12-year-old to monitor thoughts and feelings in a journal*

3 or 4 Moderate
- ➢ Intervention/assessment involves tasks that may match the physical, emotional, or cognitive capacities and needs of family members, but vague or unclear
- *e.g., Th asks mom to think of consequences for TC's behavior, but since mom was unable to think of any during the session, unclear how she will identify some after the session*

5 or 6 High
- ➢ Intervention/assessment involves tasks that match the physical, emotional, or cognitive capacities and needs of family members
- *e.g., Th facilitates job or vocational training for caregiver or older adolescent who has repeatedly failed out of school*

P6S2. The intervention was explained/discussed in a manner understandable to family members.

1 or 2 Low
- ➢ Th discusses rationales and procedures in a manner not understandable to family members
- *e.g., Th uses jargon such as "titrate" without giving lay meaning*

3 or 4 Moderate
- ➢ Rationales and procedures given to family members by Th are somewhat vague or unclear
- *e.g., when Th asks TC "do you remember what I mean by triggers?", TC says "yes" but Th does not probe to confirm that he understands*

5 or 6 High
- ➢ Th discusses rationales and procedures in a manner clearly understandable to family
- *e.g., Th asks TC if she understands the word "remission;" when TC says that she does not know, Th uses simpler words until TC is able to paraphrase*

P7S1. The intervention required daily or weekly effort by family members prior to the next session.

1 or 2 Low
- ➢ Intervention does not require family to work between-sessions
- *e.g., Th recommends no homework to family members monitoring throughout session*

3 or 4 Moderate
- ➢ Intervention requires daily or weekly action by family members between-sessions, but vague and/or do not need to be done on weekly or daily basis
- *e.g., Th asks parent to "think about" consequences for TC rather than complete concrete tasks*

5 or 6 High
- ➢ Intervention requires daily or weekly action by family members between-sessions that is specific and/or concrete
- *e.g., Th reviews with TC specific, ongoing, school-related tasks to complete and reminds TC to complete this week as well*

P7S2. The intervention required daily or weekly effort by collaborating change agents *(e.g., parole officer, teacher, family friend)* when appropriate.

1 or 2 Low
- ➢ Intervention does not utilize daily or weekly services by potential collaborating agents when they are clearly needed
- *e.g., mom says that she needs help in monitoring TC, but Th does not address potential use of family members, friends, or community resources*

3 or 4 Moderate
- ➢ Intervention involves collaborating agents, but vague and/or do not need to be done on weekly or daily basis not utilized effectively, or barriers to effective use not adequately addressed
- *e.g., Th gives information on child behavior management to teacher, but plan for teacher involvement unclear*

5 or 6 High
- ➢ When indicated, intervention requires daily or weekly action by collaborating agents that is specific and concrete
- *e.g., Th reviews with caregiver plan for teacher to call caregiver when TC does not show up for class*
- ➢ No collaborating agents utilized because no evidence that any are required
- *e.g., TC is performing well at school and generally following rules at home, and both caregivers are effectively monitoring youth and delivering contingencies to TC*

P8S1. Therapist evaluated how well the intervention has been working and the completion of previous recommendations.

1 or 2 Low
- ➢ Th did not assess degree to which family carried out ongoing intervention or recommendations from previous sessions

Appendix A. *(continued)*

P1S1. Therapist understood what key factors contributed to the target problems(s) or issue(s)

- *e.g., TC says to Th "I did what you told me to last week" and Th replies "good" but no other details are mentioned*
- ➢ If it is an early treatment session, Th fails to evaluate what actions/steps family members have taken in the past to address target problems
- *e.g., in initial session, Th describes the interventions he would like mom to use to enforce rules, but does not evaluate mom's prior successful/unsuccessful strategies*

3 or 4 Moderate
- ➢ Th addresses progress with ongoing interventions or extent to which recommendations from previous session were carried out by family members. However, if previous recommendations not/partially carried out or intervention ineffective, then Th fails to adequately evaluate and addresses barriers
- *e.g., although homework completion evaluated, Th merely reassigns incomplete homework rather than evaluating barriers to completion*
- ➢ If it is an early treatment session, Th makes some effort to evaluate what actions/steps family members have taken in the past to address target problems, but vague or unclear
- *e.g., while evaluating antecedents of TC's fighting with sister, caregiver says she grounded TC, but Th does not inquire about details*

5 or 6 High
- ➢ Th addresses progress with ongoing interventions or extent to which recommendations from previous session were carried out by family members. If previous recommendations not carried out, only partially carried out, or intervention ineffective, then Th adequately evaluates and addresses barriers
- *e.g., Th evaluates how caregiver successfully implemented "house rules" with TC and encouraged her to continue with her efforts*
- ➢ If it is an early treatment session, Th evaluates what actions/steps family members have taken in the past to address target problems
- *e.g., Th evaluates how caregiver has responded previously when TC has had temper tantrums, and caregiver reveals pattern of administering timeouts inconsistently*

P8S2. Therapist evaluated the effectiveness of interventions on youth/family/system functioning (*i.e.*, how well is the intervention improving youth/family well-being).

1 or 2 Low
- ➢ Th does not assess or poorly assesses current family/youth functioning or improvement
- *e.g., although TC is involved in drug court, and drug use is intermediary goal, there is no evaluation of whether TC is currently using drugs*

3 or 4 Moderate
- ➢ Th evaluated current family/youth/system functioning only, but links to intervention vague, unclear, or unknown
- *e.g., Th evaluated how often youth arrived at school on time since previous session, but does not discuss how mom's efforts to monitor TC and correspondence with teacher may have influenced this behavior*

5 or 6 High
- ➢ Th evaluates link between intervention and current functioning
- *e.g., Th reviews and discusses with caregiver* [1] *success in implementing consequences with youth,* [2]*chart showing how often TC had tantrums over the past week, and* [3] *links between consequences and reduction in tantrums*

P8S3. Therapist tapped relevant sources to evaluate the efficacy of the intervention(s).

1 or 2 Low
- ➢ Th did not assess progress from previous session or current functioning, or data obtained only from unreliable sources
- *e.g., although TC lied to Th about drug use in previous session (and this is known/brought up in the current session), Th asked TC if he smoked pot since the previous session but did not seek data from additional sources*

3 or 4 Moderate
- ➢ Th obtains outcome information from some reliable sources, although data from additional sources is warranted
- *e.g., Th obtains verbal report of TC's drug use from mom and TC, but no urine screen conducted*

5 or 6 High
- ➢ Th obtains outcome data from nearly all relevant sources
- *e.g., Th evaluates TC drug use using urine screen, self-report, and report from mom*

P9S1. Family members are responsible for carrying out the intervention.

1 or 2 Low
- ➢ There is little evidence that treatment goals are designed by, or interventions delivered by, family members
- *e.g., Th, rather than caregiver, is primarily responsible for implementing consequences to TC, with no discussion of how to transfer responsibility to natural agents*

3 or 4 Moderate
- ➢ Some treatment goals designed by, and interventions delivered by, family members, but Th takes on some tasks that family members appear capable of performing

Appendix A. *(continued)*

P1S1. Therapist understood what key factors contributed to the target problems(s) or issue(s)	
	• *e.g., Th allows parents to lead discussion addressing TC's recent school expulsion, but interrupts often, rarely elicits feedback from caregivers, and ultimately develops interventions that seem to match the Th's agenda rather than an agenda developed collaboratively with family*
5 or 6 High	➢ Treatment goals designed by, and interventions delivered by, family members, with Th playing primarily a supportive and consultative role
	• *e.g., Th explains consequences to TC with caregiver present, elicits related feedback from caregiver, and for homework encourages caregiver to discuss further with TC*
P9S2. The treatment promoted skills and competencies that maximized the potential for long-term change.	
1 or 2 Low	➢ No evidence that concrete skills are being developed or reinforced
	• *e.g., mom says she doesn't know how she would handle falling off the wagon again and Th doesn't address possible community resources, support groups, or concrete skills to cope with this possibility*
3 or 4 Moderate	➢ Evidence that concrete skills are being developed, or reinforced, but vague or unclear, or not generalizable
	• *e.g., Th reminds mom to use the point system if TC gets home past curfew, yet mom seems hesitant about being able to carry this out once treatment ends and Th does not address barriers*
5 or 6 High	➢ Evidence that concrete, generalizable skills are being developed or reinforced
	• *e.g., Th discusses with family plans for consulting with psychiatrist for TC's meds post-treatment, and has caregiver outline steps she will take to follow-through*

Note: P = Principle, S = Subprinciple, Th = Therapist, TC = Target Child.

Appendix B. Intraclass Correlation Coefficients (ICCs) for rater dyads on the 9 multisystemic therapy (MST) principles

Variable	Raters 1 & 2 (n = 14)	Raters 1 & 3 (n = 16)	Raters 2 & 3 (n = 14)
Principle 1	0.442*	0.847***	0.653**
Principle 2	0.507*	0.335	0.672**
Principle 3	0.587*	0.263	0.576*
Principle 4	0.660**	0.582*	0.479*
Principle 5	0.386	0.347	N/A[a]
Principle 6	N/A[a]	N/A[a]	N/A[a]
Principle 7	0.091	0.482*	0.708**
Principle 8	N/A[a]	0.469*	0.552*
Principle 9	0.258	0.481*	0.264
MPA	0.549*	0.591*	0.759***

Notes: Interrater reliability estimates using the revised MST coding protocol.; Intraclass correlation coefficients using a consistency definition; Single measures are reported.
MPA = MST Principle Adherence Composite.
[a] N/A indicates zero variances in the ratings; coefficients were unable to be calculated. Principle 6 was nearly always rated as "high," as the developmental needs of the youth were rarely ignored during the MST sessions.
* Fair reliability.
** Good reliability.
*** Excellent reliability.

References

Achenbach, T. M. (1991). *Manual for the child behavior checklist and 1991 profile.* Burlington: University of Vermont, Department of Psychiatry.

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

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Baker, J. K., Haltigan, J. D., Brewster, R., Jaccard, J., & Messinger, D. (2010). Non-expert ratings of infant and parent emotion: Concordance with expert coding and relevance to early autism risk. *International Journal of Behavioral Development*, *34*, 88–95.

Bechger, T. M., Maris, G., & Hsiao, Y. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement*, *34*, 607–619.

Breitenstein, S. M., Fogg, L., Garvey, C., Hill, C., Resnick, B., & Gross, D. (2010). Measuring implementation fidelity in a community-based parenting intervention. *Nursing Research*, *59*, 158–165.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design.* Cambridge, MA: Harvard University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Chapman, J. E., & Schoenwald, S. K. (2011). Ethnic similarity, therapist adherence, and long-term multisystemic therapy outcomes. *Journal of Emotional and Behavioral Disorders*, *19*, 3–16.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284–290.

Donohue, B., Azrin, N. H., Strada, M. J., Silver, N. C., Teichner, G., & Murphy, H. (2004). Psychometric evaluation of self- and collateral timeline follow-back reports of drug and alcohol use in a sample of drug-abusing and conduct-disordered adolescents and their parents. *Psychology of Addictive Behaviors*, *18*, 184–189.

Drapeau, M. (2014). The assessment of cognitive errors using an observer-rated method. *Psychotherapy Research*, *24*, 240–249.

Elliott, D. S., Ageton, S. S., Huizinga, D., Knowles, B. A., & Canter, R. J. (1983). *The prevalence and incidence of delinquent behavior: 1976–80 (report of the National Youth Survey, project report no. 26).* Boulder, CO: Behavioral Research Institute.

Ellis, D. A., Naar-King, S., Templin, T., Frey, M. A., & Cunningham, P. B. (2007). Improving health outcomes among youth with poorly controlled type I diabetes: The role of treatment fidelity in a randomized clinical trial of multisystemic therapy. *Journal of Family Psychology*, *21*, 363–371.

Farrington, D. P., & Loeber, R. (2000). Some benefits of dichotomization in psychiatric and criminological literature. *Criminal Behavior and Mental Health*, *10*, 100–122.

Goense, P., Boendermaker, L., van Yperen, T., Stams, G., & van Laar, J. (2014). Implementation of treatment integrity procedures: An analysis of outcome studies of youth interventions targeting externalizing behavioral problems. *Journal of Psychology*, *222*, 12–21.

Gomez-Schwartz, B. (1978). Effective ingredients in psychotherapy: Prediction of outcome from process variables. *Journal of Consulting and Clinical Psychology*, *46*, 1023–1035.

Gross, D., Fogg, L., Young, M., Ridge, A., Cowell, J. M., Richardson, R., et al. (2006). The equivalence of the child behavior checklist/1½–5 across parent race/ethnicity, income level, and language. *Psychological Assessment*, *18*, 313–323.

Henggeler, S. W. (2011). Efficacy studies to large-scale transport: The development and validation of multisystemic therapy programs. *Annual Review of Clinical Psychology*, *7*, 351–381.

Henggeler, S. W., & Borduin, C. M. (1992). *Multisystemic therapy adherence scale.* Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina: Unpublished Instrument.

Henggeler, S. W., Halliday-Boykins, C. A., Cunningham, P. B., Randall, J., Shapiro, S. B., & Chapman, J. E. (2006). Juvenile drug court: Enhancing outcomes by integrating evidence-based treatments. *Journal of Consulting and Clinical Psychology*, *74*, 42–54.

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

Henggeler, S. W., Pickrel, S. G., & Brondino, M. J. (1999). Multisystemic treatment of substance abusing and dependent delinquents: Outcomes, treatment fidelity, and transportability. *Mental Health Services Research*, *1*, 171–184.

Henggeler, S. W., Schoenwald, S. K., Borduin, C. M., Rowland, M. D., & Cunningham, P. B. (2009). *Multisystemic therapy for antisocial behavior in children and adolescents* (2nd ed.). New York, NY US: Guilford Press.

Hogue, A., Henderson, C. E., Dauber, S., Barajas, P. C., Fried, A., & Liddle, H. A. (2008). Treatment adherence, competence, and outcome in individual and family therapy for adolescent behavior problems. *Journal of Consulting and Clinical Psychology*, *76*, 544–555.

Horenstein, D., Houston, B., & Holmes, D. S. (1973). Clients', therapists', and judges' evaluations of psychotherapy. *Journal of Counseling Psychology*, *20*, 149–153.

Huey, S. J. (2001). *Adherence training manual for multisystemic therapy (MST): Anchors and guidelines for coding audiotaped therapy sessions.* Department of Psychology, University of Southern California: Unpublished Instrument.

Huey, S. J., Henggeler, S. W., Brondino, M. J., & Pickrel, S. G. (2000). Mechanisms of change in multisystemic therapy: Reducing delinquent behavior through therapist adherence and improved family and peer functioning. *Journal of Consulting and Clinical Psychology*, *68*, 451–467.

Kozlowski, S. W., & Kirsch, M. P. (1987). The systematic distortion hypothesis, halo, and accuracy: An individual-level analysis. *Journal of Applied Psychology*, *72*, 252–261.

Letourneau, E. J., Sheidow, A. J., & Schoenwald, S. K. (2002). *Structure and reliability of the MST therapist adherence measure scale in a large community sample.* Charleston: Medical University of South Carolina, Family Services Research Center.

Liang, K., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Miller, W. R. (1996). *Form 90: Structured assessment for drinking and related behaviors.* Washington, DC: National Institute on Alcohol Abuse and Alcoholism.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*, 365–383.

Podsakoff, P. M., MacKenzie, S. B., Lee, J., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, *88*, 879–903.

Reio, T. R. (2010). The threat of common method variance bias to theory building. *Human Resource Development Review*, *9*, 405–411.

Rescorla, L. A., Ginzburg, S., Achenbach, T. M., Ivanova, M. Y., Almqvist, F., Begovac, I., ... Verhulst, F. C. (2013). Cross-informant agreement between parent-reported and adolescent self-reported problems in 25 societies. *Journal of Clinical Child and Adolescent Psychology*, *42*(2), 262–273.

Robbins, M. S., Feaster, D. J., Horigian, V. E., Puccinelli, M. J., Henderson, C., & Szapocznik, J. (2011). Therapist adherence in brief strategic family therapy for adolescent drug abusers. *Journal of Consulting and Clinical Psychology*, *79*, 43–53.

Schoenwald, S. K., Carter, R. E., Chapman, J. E., & Sheidow, A. J. (2008). Therapist adherence and organizational effects on change in youth behavior problems one year after multisystemic therapy. *Administration and Policy in Mental Health and Mental Health Services Research*, *35*, 379–394.

Schoenwald, S. K., Halliday-Boykins, C., & Henggeler, S. W. (2003). Client-level predictors of adherence to MST in community service settings. *Family Process*, *42*, 345–359.

Schoenwald, S. K., Henggeler, S. W., Brondino, M. J., & Rowland, M. D. (2000). Multisystemic therapy: Monitoring treatment fidelity. *Family Process*, *39*, 83–103.

Schoenwald, S. K., Sheidow, A. J., & Letourneau, E. J. (2004). Toward effective quality assurance in evidence-based practice: Links between expert consultation, therapist fidelity, and child outcomes. *Journal of Clinical Child & Adolescent Psychology*, *33*, 94–104.

Suh, C. S., Strupp, H. H., & O'Malley, S. (1986). The Vanderbilt process measures: The psychotherapy process scale (VPPS) and the negative indicators scale (VNIS). In L. S. Greenberg, & W. M. Pinsof (Eds.), *The psychotherapeutic process: A research handbook* (pp. 285–323). New York, NY US: Guilford Press.

Thornberry, T. P., & Krohn, M. D. (2000). The self-report method for measuring delinquency and crime. *Criminal Justice*, *4*, 33–83.

Waldinger, R. J., Schulz, M. S., Hauser, S. T., Allen, J. P., & Crowell, J. A. (2004). Reading others' emotions: The role of intuitive judgments in predicting marital satisfaction, quality, and stability. *Journal of Family Psychology*, *18*, 58–71.

Waldron, H. B., Slesnick, N., Brody, J. L., Turner, C. W., & Peterson, T. R. (2001). Treatment outcomes for adolescent substance abuse at four- and seven-month assessments. *Journal of Consulting and Clinical Psychology*, *69*, 802–813.

Weiss, B., Han, S., Harris, V., Catron, T., Ngo, V. K., Caron, A., Gallop, R., & Guth, C. (2013). An independent randomized clinical trial of multisystemic therapy with non–court–referred adolescents with serious conduct problems. *Journal of Consulting and Clinical Psychology*, *81*, 1027–1039.