Published in final edited form as:
J Child Adolesc Subst Abuse. 2008 June 1; 17(3): 47–68. doi:10.1080/15470650802071655.

Developing a Measure of Therapist Adherence to Contingency Management: An Application of the Many-Facet Rasch Model

Jason E. Chapman, **Ashli J. Sheidow**, **Scott W. Henggeler**, **Colleen Halliday-Boykins**, and **Phillippe B. Cunningham**
Family Services Research Center, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina.

Abstract

A unique application of the Many-Facet Rasch Model (MFRM) is introduced as the preferred method for evaluating the psychometric properties of a measure of therapist adherence to Contingency Management (CM) treatment of adolescent substance use. The utility of psychometric methods based in Classical Test Theory was limited by complexities of the data, including: (a) ratings provided by multiple informants (i.e., youth, caregivers, and therapists), (b) data from separate research studies, (c) repeated measurements, (d) multiple versions of the questionnaire, and (e) missing data. Two dimensions of CM adherence were supported: adherence to Cognitive Behavioral components and adherence to Monitoring components. The rating scale performed differently for items in these subscales, and of 11 items evaluated, eight were found to perform well. The MFRM is presented as a highly flexible approach that can be used to overcome the limitations of traditional methods in the development of adherence measures for evidence-based practices.

Keywords

Contingency Management; therapist adherence; Rasch model

The Contingency Management (CM) treatment model developed by Azrin and his colleagues (Donohue & Azrin, 2001) has shown promise in treating adolescent substance use in conjunction with family therapy. Results from randomized trials demonstrated significantly better outcomes for youths in the CM conditions, compared with supportive counseling, for drug use abstinence, mental health and conduct problems, and employment/school attendance (Azrin, Acierno et al., 1996; Azrin, Donohue et al., 1994). To ensure the integrity of CM implementation in these clinical trials, Azrin and colleagues provided intensive and ongoing oversight of the therapy sessions. This oversight included (1) extensive training, modeling, and role-playing with corrective feedback to therapists; (2) ongoing audiotape coding with corrective feedback provided to therapists; (3) observation of therapy sessions with corrective feedback; (4) detailed session checklists prompting use of techniques, reviewed weekly with therapists; and (5) weekly written documentation by therapists of techniques used, youth and family participation, and progress toward treatment goals. Notably, fidelity data from session checklists and ratings of audiotapes showed greater than 95% adherence (Azrin, Donohue et al., 1994).

Correspondence concerning this article should be addressed to Jason E. Chapman, Family Services Research Center, Department of Psychiatry and Behavioral Sciences, Medical University of South Carolina, 67 President Street, Suite CPP, P.O. Box 250861, Charleston, South Carolina, 29425. Electronic mail may be sent to chapmaja@musc.edu.

Adherence to treatment protocols, however, can fall precipitously when treatments are transported to community-based practitioners (e.g., Henggeler, Melton, Brondino, Scherer, & Hanley, 1997). Further, this decline in fidelity has been linked empirically to lowered effectiveness of the intervention. For example, several studies have supported linkages between adherence to Multisystemic Therapy (MST) treatment principles and clinical outcomes (e.g., Henggeler, et al., 1997; Henggeler, Pickrel, & Brondino, 1999; Schoenwald, Sheidow, Letourneau, & Liao, 2003). Likewise, favorable outcomes for other evidence-based practices (e.g., Functional Family Therapy, Program for Assertive Community Treatment) have been attenuated by low therapist fidelity to the corresponding treatment protocols (Weisz, Weersing, & Henggeler, 2005).

If achieving desired clinical outcomes relies, in part, on the fidelity of therapist implementation of the treatment model, then attempts to transport evidence-based practices to community settings requires effective measures of treatment fidelity. Although the observational methods used for monitoring and measuring fidelity in the CM trials noted previously were highly effective, they also were very time consuming and, as such, expensive. Indeed, most of the research-supported treatments of adolescent substance abuse rely on videotape assessments of therapist behavior to assess treatment fidelity (e.g., Liddle et al., 2001; Santisteban et al., 2003). Although observational methods are likely the most accurate way to assess treatment fidelity and can be used effectively in highly resourced university-based treatment research, they are unwieldy for the large-scale transport of evidence-based practices to community settings (e.g., across hundreds of practitioners across a state). The large-scale transport of evidence-based practices requires methods for evaluating therapist fidelity that can be employed reliably with fewer resources and incorporated effectively into a continuous quality improvement system. Toward this end, Henggeler and colleagues developed a brief paper-and-pencil measure of therapist adherence to CM that has been used in three research studies described subsequently. This questionnaire is administered to therapists, caregivers, and youths on a monthly basis to evaluate therapist behavior. Evaluating the psychometric properties of this measure, however, has presented numerous challenges that could not be addressed by traditional methods of scale development.

The construction of a brief, psychometrically sound, multi-respondent measure of therapist adherence to CM is associated with several unique challenges. First, although the multiple respondents (e.g., youths, caregivers, therapists) report on the behavior of a single, common therapist, they might use each questionnaire item in a different manner. Second, the inherent nesting of data can be associated with variability in adherence scores (e.g., the scores provided by youths who have the same therapist are likely more similar than the scores provided by youths who have different therapists). Third, the questionnaires often are administered repeatedly, introducing the potential for items to perform differently over time. Fourth, missing data are inevitable when large numbers of cases are assessed repeatedly through multiple respondents. Finally, the application of the questionnaire in varied settings can result in the addition, deletion, and/or modification of items (e.g., investigators in different states can make unilateral decisions about the inclusion and exclusion of questionnaire items). Flexible and comprehensive psychometric methods are necessary to overcome these challenges to the development of a measure of CM adherence.

A primary purpose of this article is to describe and demonstrate how the Rasch model, in contrast with traditional psychometric methods based in Classical Test Theory (CTT; i.e., True Score Theory), is ideally suited to address the aforementioned measurement development challenges. The cardinal difference between CTT and the Rasch model is that, in CTT, items and respondents are inherently confounded. That is, information about the functioning of items (e.g., item difficulty) is intertwined with information about the performance of respondents (e.g., person ability). CTT methods do not provide a method for separating the two, whereas

Rasch modeling delineates item and respondent performances distinctly. One of the greatest problems posed by the confounding of items and respondents pertains to varying “difficulty” or “severity” of items within a measure (Bond & Fox, 2001; Wright & Mok, 2000). The term “item difficulty” is readily exemplified in the context of achievement testing. For instance, two students might get the same number of SAT items correct, but one student correctly answered the easiest items and the other student answered the most difficult items. Thus, the students’ scores should be based on their performance relative to where items fall on the spectrum of difficulty, resulting in a higher score for the student who correctly answered more difficult items. Developers of standardized achievement tests have long been utilizing the family of Item Response methods, of which the Rasch model is a special case, for this and other reasons.

More specific to the area of adolescent substance abuse, consider a brief substance use questionnaire in which an adolescent endorses “yes” or “no” for use of different substances during the previous 30 days. If the count of positive responses is used to indicate severity, then two youths could receive the same quantitative score despite dramatic differences in the quality of their substance use. For example, a score of “2” could be obtained for endorsement of marijuana and alcohol use, but also for endorsement of heroin and cocaine use. In this example, alcohol and marijuana might be considered “less severe” (or easier to endorse) items on the spectrum, while heroin and cocaine might be considered “more severe” (or more difficult to endorse). Extending this issue to a questionnaire measuring therapist adherence, two therapists might receive the same adherence score, even though one therapist adhered to the easier components of the treatment and the other therapist adhered to the most difficult components.

Thus, the underlying aim of the questionnaire (e.g., the measurement of achievement, substance use severity, or therapist adherence) can be confounded by the characteristics of the items. The reverse also is true. That is, when the aim of the assessment is evaluation of items (e.g., psychometric questions), this information can be confounded by the characteristics of the sample of respondents (e.g., items completed with reference to a group of non-adherent therapists will appear to be much more difficult than the same group of items completed with reference to a highly adherent group of therapists). The Rasch model overcomes the inherent confounding of items and persons by providing separate information about people (i.e., person ability), free of the influence of the sample of items, and likewise, information about items (i.e., item difficulty), free of the influence of the sample of people. Rather than ignoring item characteristics or applying an arbitrary adjustment to account for varying characteristics of items, the Rasch model allows for item and person characteristics to be evaluated and estimated separately.

The Rasch model can address several other limiting features of CTT as well. For instance, the Rasch model can evaluate rating scale performance to ensure that the response options function as intended. CTT methods implicitly assume that the response scale for each item (e.g., 5-point Likert scale) functions as intended, but offers no techniques for evaluating this assumption. Violation of this fundamental assumption could result in a score that misrepresents the underlying construct the questionnaire (or subscale) serves to quantify. For instance, rating scales are intended to function such that each point on the response continuum is meaningful and distinct from the other options. While test developers endeavor to achieve this aim, the resulting scales often do not function as intended (Lopez, 1996). As an illustration, respondents might struggle to discriminate between item options such as “almost never” and “sometimes,” underutilizing a given option or using the options in an inconsistent manner. Rating scales also might function differently from item to item on the same questionnaire. Such discrepancies can be identified and accurately adjusted with the Rasch model, producing true interval scaled data (using the log-odds unit, or logit; Wright, 1993). This feature is particularly important both because CTT methods assume observations are on an interval scale and the vast majority of mathematical and statistical operations that are used to analyze questionnaire data require interval-level scaling.

Another limitation addressed by the Rasch model is that traditional scale development methods often do not map onto the realities of data collection. In particular, CTT methods typically require complete data, listwise deletion of cases with missing data, or the imputation of missing responses. This is not the case with the Rasch model, as it includes missing data in analyses. This capability is important given that researchers, particularly services researchers conducting projects with “real world” clinicians and agencies, commonly are faced with at least some missed assessments or missing informants on a multi-informant measure. Also, data can be missing due to the use of alternate forms of a questionnaire (e.g., to prevent effects related to repeated administration of items, resulting from the addition, deletion, and/or modification of items). The Rasch model overcomes these barriers by utilizing the information from alternate forms and providing results that are directly comparable through “test equating” methods (Wolfe, 2000).

Moreover, when data are nested and cross-classified, the standard Rasch model can be expanded to a Many-Facet Rasch Model (MFRM; Smith et al., 2002). The MFRM traditionally is used when ratings of a specific person are provided by multiple “judges” (Linacre & Wright, 2002), whereby a “facet” to indicate which judge provided the rating is included in the model. Extending this concept to other characteristics of a testing situation can generate a viable method for conducting accurate scale development. For instance, instead of representing judges, facets in the model can represent multiple informants (e.g., youth, caregiver, therapist), time (i.e., repeated administrations), common targets (e.g., nesting within the same therapist), or other characteristics (e.g., data gathered from multiple studies). As an example, informants might differ substantially on the severity of their ratings. That is, therapists have detailed knowledge of the treatment model they are aiming to provide families and so might be harsher judges of their adherence behavior compared to caregivers or youth ratings of that same behavior. Similarly, caregivers might be more observant of therapist behavior than are youth. The MFRM allows the evaluator both to assess and control for these confounds by putting therapists, caregivers, and youth on the same “yardstick,” adjusting for the influence of the other facets. Hence, the Rasch model represents an optimal solution to several limitations encountered by researchers attempting scale development with complex data.

The present article has two overarching goals. The first objective is to describe the psychometric properties of a newly developed measure of therapist adherence to Contingency Management in the treatment of adolescent substance abuse. The Contingency Management Therapist Adherence Measure (CM-TAM) requires relatively few resources to implement and adds comparatively little burden to “real world” agencies, therapists, and families. This aim is accomplished using data from three independent research studies that utilized the CM-TAM. The second objective is to illustrate the application of the Many-Facet Rasch Model to address scale development problems posed by alternate versions of the CM-TAM, repeated administrations, and multiple types of respondents. Here, a detailed overview of the key features of Rasch modeling, including data considerations, calibration, model refinement, and interpretation is provided.

Method

Project Data Sets Included

The present study leveraged data from three studies that recently have been completed.

Juvenile Drug Court Integration—The first study was a randomized trial evaluating the effectiveness of juvenile drug court, the effects of integrating an evidence-based treatment (i.e., MST) into the drug court process, and the effects of integrating CM techniques into the MST treatment protocol for juvenile substance abuse outcomes (Henggeler et al., 2006). Participants in the study were juvenile offenders meeting diagnostic criteria for substance abuse or dependence and their caregivers. Families were randomized into four treatment conditions. The CM-TAM was completed by 279 respondents (123 youths, 130 caregivers, 26 therapists) providing 1,200 CM-TAM reports over an average of 3.3 administrations (*SD* = 1.50).

Transportability within an Existing Evidence-Based Practice—In the second study, quality assurance strategies to sustain treatment fidelity to CM were compared for substance abusing delinquents within the context of an existing evidence-based practice, MST (Sheidow, Henggeler, Cunningham, Donohue, Ford, & Shapiro, 2006). Following a 2-day CM workshop, MST teams were randomized to CM intensive quality assurance versus no sustained quality assurance. CM-TAM reports were collected monthly from youths, caregivers, and therapists. This resulted in 639 CM-TAM reports provided by 193 respondents (79 youths, 78 caregivers, 36 therapists) over an average of 3.1 administrations (*SD* = 1.40).

Statewide Transportability of CM to Mental Health and Substance Abuse Practitioners—The third study evaluated practitioner demographic, professional training, organizational, and service sector predictors of voluntary attendance at a CM workshop and implementation of CM in the 6 months following the workshop (Henggeler et al., 2006). Participants were therapists and supervisors in substance abuse (*N* = 178 across 30 agencies) and mental health provider organizations (*N* = 365 across 14 agencies). Therapists provided self-reports on a maximum of three substance abusing youth clients during each monthly assessment. This resulted in 1,790 CM-TAM self-reports provided by 224 therapists over an average of 4.4 occasions (*SD* = 1.86).

Measure

A total pool of eleven items was used to assess adherence in CM across the three studies (Table 1). Each study used a different 9-item version of the CM Therapist Adherence Measure (CM-TAM). Seven items were administered across all studies, and the two variable items were the result of study-specific adaptations. Thus, ratings are available on a total of five items targeting the Cognitive-Behavioral (CB) techniques and six items targeting the Monitoring (MON) techniques that comprise CM. Items were rated according to a 5-point scale with response options of 1 = Not at All, 2 = A Little, 3 = Some, 4 = Pretty Much, 5 = Very Much.

Data Analysis

Combined Data—The data used in the present analyses were comprised of 3,629 CM-TAM administrations (*M* = 3.6 administrations per respondent, *SD* = 1.69) of the 11 CM-TAM items nested within 696 respondents (202 youths, 208 caregivers, 286 therapists). Respondents rated the behavior of (were nested within) 285 different therapists.

Many-Facet Rasch Model (MFRM)—The standard Rasch model is comprised of two facets, item difficulty and person ability, with dichotomous responses (Rasch, 1966). According to this model, the probability of a respondent endorsing a given item is the net result of the interaction between the ability of the respondent and the difficulty of the item (Wright & Mok, 2000). The dichotomous Rasch model is readily extended to a Rasch rating scale model through the inclusion of a parameter that represents the probability of transitioning from one rating scale category to the next.

As described by Linacre (1994), the MFRM further extends the standard two facet rating scale model. For MFRM, the probability of a given CM-TAM response is the result of the therapist's level of CM adherence, the difficulty of the adherence component, and the leniency of the respondent (i.e., the tendency of the respondent to be a harsh or lenient judge by providing ratings that are systematically low or systematically high, respectively; Linacre & Wright, 2002). The MFRM was used to evaluate the functioning of the CM-TAM items (facet 1), adjusting for the sampling effects of the respondent's leniency (facet 2), the adherence level of the target therapist (facet 3), the leniency of the respondent type (facet 4; i.e., the tendency of youth, caregivers, and therapists to provide high/low ratings), the adherence level of associated study (facet 5), and the leniency of reports at a given time (facet 6). The MFRM was calibrated using FACETS computer software (Linacre, 2004a).

Steps of MFRM Analyses and Interpretive Guidelines—The MFRM includes several critical steps.

Dimensionality: A fundamental assumption of the Rasch model is unidimensionality, indicating that the items under investigation measure a single attribute (Bond & Fox, 2001). The CM-TAM items, however, were designed to measure adherence to the conceptually distinct Monitoring (MON) and Cognitive Behavioral (CB) components of CM. As a result, CM-TAM dimensionality was evaluated using three sources of information. First, Principal Components Analysis (PCA) of standardized Rasch residuals was used to extract the factor explaining the greatest percentage of residual variance (Linacre, 1998). Unidimensionality is supported if this variance represents random noise. However, if a factor explains a “non-trivial” portion of the variance, performing a Rasch calibration on each group of items (i.e., dimensions) is warranted. WINSTEPS (Linacre, 2004b) was used to perform the PCA of Rasch residuals. The critical eigenvalue size for non-trivial dimensionality was determined by the conventional cut-off of 1.4 (Smith & Miao, 1994) and the parallel analysis eigenvalue of 1.1 (i.e., the 95th percentile eigenvalue from 100 iterations of random data with the same number of variables and cases as the present data; see Raiche [2005] for details of this method). Second, the item fit statistics, described subsequently, provided another indicator of dimensionality (Smith, 1996) and were examined for groups of items departing significantly from their expected values. Third, the Rasch person-measures (i.e., scores) derived from separate calibration of the CB and MON items were cross-plotted. This illustrated the degree to which similar conclusions about a therapist's level of adherence would be obtained using either group of items, as would be expected if the items measured the same dimension. Plots departing from a straight line suggest that groups of items measure different dimensions (Linacre, 1998, 2004b).

Rating scale functioning: FACETS (Linacre, 2004a) provides several indicators of adequacy of rating scale performance. Six indicators were evaluated in the present analyses, as detailed by Linacre (2002) and Bond and Fox (2001). First, “category use statistics,” or the frequency of responses in each category, were assessed for a consistent distribution across rating categories. A recommended guideline is at least 10 observed responses per category. Second, the average Rasch respondent leniency estimates for those who endorsed a given response category were examined to assess the degree to which higher category utilization was associated with increasing respondent leniency. Third, threshold estimates, indicating the point at which the probability of endorsement of two adjacent categories is equal, should be spaced by at least 1.0 logit, indicating that each rating scale transition is a distinct point on the rating scale continuum. Fourth, step calibrations were evaluated to determine difficulty of selecting one response category over another and should increase as the response category increases. Fifth, category fit statistics were examined as an indication of the degree to which categories performed as predicted. Standardized OUTFIT values exceeding 2.0 indicate that the category contributed more “noise” than precision to the data. Sixth, the response probability curves provided an illustration of the statistics described above.

Item fit: FACETS provides several statistics for both items and respondents that quantify the degree to which the observed data fit the expected model. Fit statistics are based on the differences between the observed and expected response for each person on each item (Bond & Fox, 2001). Thus, small residual values indicate that the observed response was close to the expectation, and large residual values indicate that the response was unexpected. The ZSTD OUTFIT value has been found to perform most effectively for identifying misfitting items (Smith, 2000; Smith, Schumaker, & Bush, 1998). The critical scores for the ZSTD OUTFIT statistics are ±2, with values ≤ −2, suggestive of redundancy, which is less of a concern when the primary aim of the analyses is not item reduction.

Separation reliability: FACETS also provides separation reliability estimates for each facet in the model. Separation reliability refers to the number of levels of a given facet reliably differentiated by the other facets in the model (Smith, 2001). In the present context, the item separation reliability indicates the number of levels of therapist adherence defined by the items (Bond & Fox, 2001).

Results

Dimensionality

PCA of Rasch Residuals—The PCA of Rasch residuals revealed an eigenvalue of 3.4 for the first factor, accounting for 9.6% of the residual variance. The eigenvalue magnitude exceeded both the conventional and parallel analysis cutoffs, thereby rejecting the unidimensionality assumption. Rather, as expected, the residuals suggested two separate factors. All CB items correlated positively with the first factor (range = .54 to .74) and all MON items correlated negatively with this factor (range = −.18 to −.63).

Item Fit—CB and MON items were significantly misfitting, but in opposite ways. This model was also re-calibrated using FACETS to evaluate item fit when adjusting for the other facets under investigation. The results were consistent, that is, OUTFIT ZSTD values for all CB items were < −2.0 and ≥ 2.0 for all MON items.

Cross-plot of Respondent Measures—Respondent measures were calibrated separately based on CB and MON items, and the measures were cross-plotted. The resulting plot departed markedly from a straight line. Because the scores were minimally correlated, this indicates that the MON and CM score would lead to different conclusions about a given therapist's level of CM adherence. If the items measured the same dimension, scores based on the two groups of items would lead to the same conclusion about the level of adherence.

Summary—As expected based on the measure's design, each source of evidence indicated that the CB and MON form separate dimensions. Thus, CB and MON adherence items were calibrated separately in the MFRM.

Rating Scale Functioning

Category Use Statistics—The category use statistics for the 5-point rating scale for CB items revealed that the percentage endorsement for response options 1–5 were 14%, 7%, 15%, 25%, and 39%, respectively. A similar pattern was revealed for MON items. The percentage endorsement for response options 1–5 were 28%, 6%, 10%, 17%, 39%, respectively. Thus, for CB items, the response option of 2 ("*A Little*") was underutilized, and for the MON items, options 2 ("*A Little*") and 3 ("*Some*") were underutilized.

Average Respondent Measures—For CB and MON items, the average and expected respondent measures (i.e., "scores") increased with the corresponding response options. Thus, as expected, higher ratings were associated with higher levels of adherence.

Threshold Estimates—The distance between adjacent threshold estimates should be at least 1.0 logit. For CB and MON items, this was not observed. The distances between the five response categories for CB and MON items were .69 (between 1 and 2), .31 (2–3), .45 (3–4), .83 (4–5) and .31 (between 1 and 2), .14 (2–3), .19 (3–4), .34 (4–5), respectively.

Step Calibrations—Disordering of step calibrations for CB and MON items was present, as illustrated for the CB dimension in Figure 1. The curves for each response option represent the predicted probability of a given rating at each point along the x-axis. Each curve should form a "peak" above the other curves at some point along this continuum. If the curve does not form a peak, it indicates that the response option was not readily or accurately distinguished from the adjacent options. As illustrated for CB items, the curves for ratings of "2" and "3" did not form a peak above the other response options. For MON items, ratings of "2," "3," and "4" did not emerge as peaks.

Category Fit Statistics—The OUTFIT mean-square value for each response category was within the accepted bounds indicating that the response options were not used in a "noisy" or unpredictable manner.

Summary—The results of the Rasch rating scale analysis provided evidence that the 5-point rating scale did not function as intended for CB or MON items. The middle response options were underutilized for both CB and MON items, and there was limited differentiation between adjacent response categories. Thus, prior to evaluation of item fit, the rating scale categories were adjusted according to the methods described by Linacre (2002), as described next.

Rating Scale Optimization—Inspection of the response probability curves illustrated that the 5-point scale functioned as a 3-point scale for CB items (Figure 1), and as a 2-point scale for MON items. The guidelines detailed previously and the conceptual meanings of the response options (i.e., wording of the response options appeared to overlap in meaning) were used to guide the evaluation of alternative groupings. For CB items, combining the three middle rating scale categories (i.e., ratings of 2 = "*A Little*", 3 = "*Some*", 4 = "*Pretty Much*") produced a well-functioning three point scale with category descriptors of "*None*" (rating of 1), "*Some*" (ratings of 2, 3, 4), and "*Very Much*" (rating of 5). The optimized CB probability curves are presented in Figure 2. For the MON items, combining rating categories of 1, 2, 3, 4 produced a well-functioning dichotomous scale with category descriptors of "*No*" (ratings of 1, 2, 3, 4) and "*Yes*" (rating of 5). This approach to rescoring CB and MON items was applied in subsequent models.

Item Fit and Separation Reliability

Cognitive Behavioral—Three of the five CB items were found to fit the model adequately according to the OUTFIT ZSTD statistics. The OUTFIT value for item 3 was −2.6, indicating that the item content is possibly redundant. However, this item was retained as test length was not a concern. The OUTFIT value for CB item 7 was 5.2, indicating that it contributed more noise than precision to the measurement of CB adherence. Also of note, the CB items were of similar difficulty, indicating that the items covered a small portion of the range of therapist CB adherence (see Figure 3). The CB item separation reliability was 4.6, indicating that the items were able to reliably distinguish approximately 4 statistically distinct levels of therapist adherence to the cognitive behavioral aspects of CM. Item 7 was removed and the model was recalibrated. The remaining items were not significantly misfitting, and the OUTFIT value for CB item 3 dropped to −2.0, indicating that its fit to the model improved.

Monitoring—Three of the MON items were found to fit the model adequately. OUTFIT ZSTD for items 5 and 9 were −3.1 and −4.0, indicating possibly redundant content. Item 6 was significantly misfitting, with an OUTFIT value of 5.7. The items covered a 2.30 logit range of therapist MON CM-adherence. The most difficult MON CM items to endorse were 10 and 11; items 4, 5, and 9 were of average difficulty; and item 6 was the easiest to endorse. The MON CM item separation reliability was 12.7, indicating that the items were able to reliably distinguish approximately 12 statistically distinct levels of therapist adherence to the monitoring aspects of CM. Item 6 was removed and the model was recalibrated. MON 4 emerged as significantly misfitting (OUTFIT = 4.7). This item was removed, and the remaining MON items fit the model well.

Facets—Illustration of each facet in the model is provided in Figure 3 for the CB dimension. Across both CB and MON, *Respondents* varied widely in level of leniency. *Targets* were less variable, but still covered a wide range of therapist adherence for CB and covered a moderate range for MON. For CB, the *Respondent Type* facet showed that caregivers, youths, and therapists were similar in level of leniency, with caregivers and youths being slightly more lenient than therapists, as expected. For MON, the *Respondent Type* facet showed that youths and therapists were similar in level of leniency, but less lenient than caregivers. Across both CB and MON, *Studies* were similar in adherence level, with the Transportability within an Existing Evidence-Based Practice study having slightly lower adherence levels than the other two studies. Finally, the *Time* facet was largely invariant for both CB and MON dimensions, with similar levels of ratings across administrations.

Discussion

This study evaluated the psychometric properties of a recently developed, brief measure of therapist adherence to CM for the treatment of adolescent substance abuse. Complexities of the data, including repeated measurements and multiple levels of nesting, limited the viability of traditional psychometric methods based in Classical Test Theory. Instead, a unique application of the Many-Facet Rasch Model was utilized, providing a flexible approach with a number of strengths.

The results supported the presence of two distinct dimensions of adherence to CM: Cognitive-Behavioral and Monitoring techniques. The original rating scale was found to perform differently for CB and MON items. That is, respondents distinguished a middle level of CB adherence, but considered MON adherence as either having occurred or not. These findings are consistent with the discrete nature of the MON components (e.g., performing a drug test). Additionally, the middle category (“*Some*”) for CB items covered a wide range of the rating scale construct, suggesting that in future applications of the CM-TAM, an additional level might be discriminated. Future applications of the CM-TAM should revise the response options based on these findings.

One CB item and two MON items were found to perform poorly (items 4, 6, and 7). Future applications of the CM-TAM, should likely remove or modify these items. Though findings support the performance of the remaining CB and MON items, the limited range of therapist adherence targeted by the items is noteworthy. Thus, a key recommendation for future use of the CM-TAM is the development of additional items that are designed specifically to target observed gaps in the scale (i.e., the highest and lowest levels of adherence) using methods described by Stone (2003). For example, preliminary implementation data have suggested that the therapists rarely develop maintenance plans with families. Thus, a “difficult” item could be developed around this aspect of CM adherence that would target the most highly adherent therapists. Finally, there was some indication that therapist reports might be more conservative than the reports of caregivers and/or youth. Future applications of the CM-TAM should continue to evaluate the impact of different types of respondents.

The aim of the study focused on item performance; thus, less attention was given to the other facets in the model. It would be useful, however, to further explore the results provided for each of these facets. Also, the presence of “disjoint subsets” (Linacre, 2004a) was a limitation of the data. For the optimal calibration of items using FACETS, each person providing ratings should rate multiple therapists. However, this cannot be the case for caregiver and youth ratings of therapist behavior. As a result, there is ambiguity in the calibration of the model. Although several possible remedies to this problem are feasible, the ideal solution is the use of the Multilevel Measurement Model (Beretvas & Kamata, 2005). This recent extension of the Rasch model directly accounts for the nesting of data.

Because the ultimate value of a fidelity measure is its ability to predict outcomes, future investigations should evaluate the associations between the CM adherence components and key outcomes. In the context of the present study, this can be accomplished by utilizing the validated CB and MON components to predict drug use, criminal charge, and behavioral outcomes within the three CM studies previously described. To further facilitate the successful transport of evidence-based practices into community-based settings, future work should also continue to develop and evaluate economically viable means of assessing adherence. The present study illustrates a procedure for evaluating and improving such measures using a highly flexible and sophisticated psychometric approach.

Acknowledgments

This manuscript was supported by grants K23DA015658, R01DA015844, R01DA013066, R01DA17487, and R01ZDA from the National Institute on Drug Abuse; R01AA122202 from the National Institute on Alcoholism and Alcohol Abuse and SAMHSA/Center for Substance Abuse Treatment; and P01HS10871 from the Agency for Healthcare Research & Quality. The authors gratefully acknowledge the support of Dr. David Henry in developing this manuscript.

References

Azrin NH, Acierno R, Kogan ES, Donohue B, Besalel VA, McMahon PT. Follow-up results of supportive versus behavioral therapy for illicit drug use. Behaviour Research & Therapy 1996;34:41–46. [PubMed: 8561763]

Azrin NH, Donohue B, Besalel VA, Kogan ES, Acierno R. Youth drug abuse treatment: A controlled outcome study. Journal of Child & Adolescent Substance Abuse 1994;3:1–16.

Azrin NH, McMahon PT, Donohue B, Besalel VA, Lapinski KJ, Kogan ES, et al. Behavior therapy for drug abuse: A controlled treatment outcome study. Behaviour Research & Therapy 1994;32:857–866. [PubMed: 7993330]

Bond, TG.; Fox, CM. Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Erlbaum; 2001.

Donohue, B.; Azrin, NH. Family behavior therapy. In: Wegner, EF.; Waldron, HB., editors. Innovations in adolescent substance abuse interventions. New York: Pergamon Press; 2001. p. 204-227.

Henggeler SW, Chapman JE, Rowland MD, Halliday-Boykins CA, Randall J, Shackelford J, et al. If you build it, they will come: A statewide study of practitioner and organizational willingness to learn about an evidence-based practice for adolescent substance abuse. 2006 Manuscript submitted for publication.

Henggeler SW, Halliday-Boykins CA, Cunningham PB, Randall J, Shapiro SB, Chapman JE. Juvenile drug court: Enhancing outcomes by integrating evidence-based treatments. Journal of Consulting and Clinical Psychology 2006;74:42–54. [PubMed: 16551142]

Henggeler SW, Melton GB, Brondino MJ, Scherer DG, Hanley JH. Multisystemic therapy with violent and chronic juvenile offenders and their families: The role of treatment fidelity in successful dissemination. Journal of Consulting & Clinical Psychology 1997;65:821–833. [PubMed: 9337501]

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

Liddle HA, Dakof GA, Parker K, Diamond GS, Barrett K, Tejeda M. Multidimensional family therapy for adolescent drug abuse: Results of a randomized clinical trial. American Journal of Drug & Alcohol Abuse 2001;27:651–688. [PubMed: 11727882]

Linacre, JM. Many-facet Rasch measurement. Chicago: MESA Press; 1994.

Linacre JM. Structure in Rasch residuals: Why principal components analysis? Rasch Measurement Transactions 1998;12:636.

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

Linacre, JM. Facets Rasch measurement computer program. Chicago: 2004a. Winsteps.com

Linacre, JM. WINSTEPS Rasch measurement computer program. Chicago: 2004b. Winsteps.com

Linacre JM, Wright BD. Construction of measures from many-facet data. Journal of Applied Measurement 2002;3:486–512. [PubMed: 12486312]

Lopez W. Communication validity and rating scales. Rasch Measurement Transactions 1996;10:482.

Raiche G. Critical eigenvalue sizes in standardized residual principal components analysis. Rasch Measurement Transactions 2005;19:1012.

Rasch G. An item analysis that takes individual differences into account. The British Journal of Mathematical and Statistical Psychology 1966;19:49–57. [PubMed: 5939145]

Santisteban DA, Coatworth JD, Perez-Vidal A, Kurtines WM, Schwartz SJ, LaPerriere A, Szapocznik J. Efficacy of brief strategic family therapy in modifying Hispanic adolescent behavior problems and substance use. Journal of Family Psychology 2003;17:121–133. [PubMed: 12666468]

Schoenwald SK, Sheidow AJ, Letourneau EJ, Liao JG. Transportability of evidence-based treatments: Evidence for multi-level influences. Mental Health Services Research 2003;5:223–239. [PubMed: 14672501]

Sheidow, AJ.; Henggeler, SW.; Cunningham, PB.; Donohue, BC.; Ford, JD.; Shapiro, SB. Transporting Contingency Management for Youth Treated with Multisystemic Therapy. In: Schoenwald, SK., (Chair), editor. Key Findings in the Transport and Implementation of Evidence-Based Treatments to Community Settings; Symposium conducted at the 2006 Joint Meeting on Adolescent Treatment Effectiveness; Baltimore, MD. 2006 Mar.

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 2002b;3:205–231. [PubMed: 12011501]

Smith EV Jr, Conrad KM, Chang K, Piazza J. An introduction to Rasch measurement for scale development and person assessment. Journal of Nursing Measurement 2002;10:189–206. [PubMed: 12885145]

Smith RM. A comparison of methods for determining dimensionality in Rasch measurement. Structural Equation Modeling 1996;3:25–40.

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

Smith RM, Schumacker RE, Bush JJ. Using item mean squares to evaluate fit to the Rasch model. Journal of Outcome Measurement 1998;2:66–78. [PubMed: 9661732]

Smith, RM.; Schumacker, RE.; Bush, JJ. Examining replication effects in Rasch fit statistics. In: Wilson, M.; Engelhard, G., editors. Objective measurement: Theory into practice. Vol. Vol. 5. Stamford, CT: Ablex Publishing Corp.; 2000. p. 303-317.

Stone MH. Substantive scale construction. Journal of Applied Measurement 2003;4:282–297. [PubMed: 12904678]

Weisz JR, Weersing VR, Henggeler SW. Jousting with straw men: Comment on Westen, Novonty, and Thompson-Brenner (2004). Psychological Bulletin 2005;131:418–426. [PubMed: 15869338]

Wolfe EW. Equating and item banking with the Rasch model. Journal of Applied Measurement 2000;1:409–434. [PubMed: 12077465]

Wright BD. “Logits?”. Rasch Measurement Transactions 1993;7:288.

Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Archives of Physical Medicine and Rehabilitation 1989;70:857–860. [PubMed: 2818162]

Wright, BD.; Masters, GN. Rating scale analysis. Chicago: MESA Press; 1982.

Wright BD, Mok M. Rasch models overview. Journal of Applied Measurement 2000;1:83–106. [PubMed: 12023559]

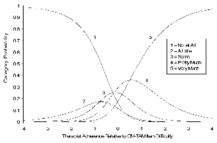

Figure 1.

CB response probability curves.

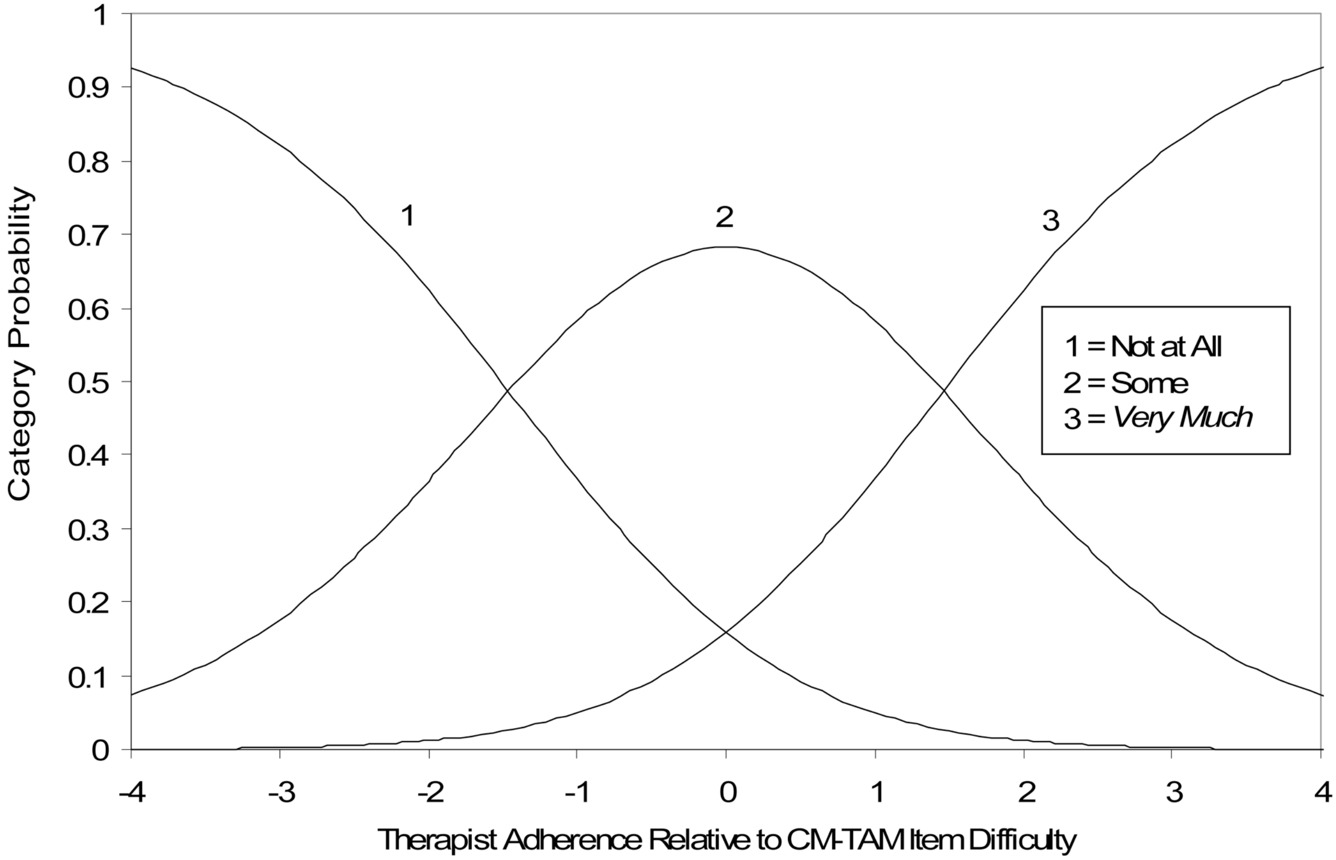

Figure 2.
CB optimized response probability curves.

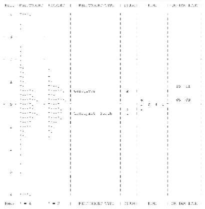

Figure 3.
Variable map of the six facet model.

Table 1

CM-TAM Items Administered to Respondents by Study

		Study[b]		
	CM-TAM Items[a]	**1**	**2**	**3**
1.	I helped the child think of ways to tell people that he/she does not want to use drugs	✓	✓	✓
2.	I helped the child practice what to do when things or triggers happen that might cause him/her to use drugs or alcohol	✓	✓	✓
3.	I helped the child come up with ways to get out of situations that involve drug use		✓	✓
4.	I gave negative consequences (punishments) to the child if the drug screen was dirty and positive consequences (rewards) if the screen was clean	✓	✓	✓
5.	I informed family members of the child's drug test results within 24 hours	✓	✓	
6.	I tested the child for alcohol or drug use by breathalyzer or drug screen	✓	✓	✓
7.	I helped the child make a list of things or triggers that might cause him/her to use drugs or alcohol	✓	✓	✓
8.	I helped the child practice how to act if someone offers him/her drugs, that is, ways to refuse drugs	✓	✓	✓
9.	I made sure the caregiver gave a positive or negative consequence for the child's drug screen results	✓	✓	✓
10.	I tried to get family members to be in charge of collecting the child’s drug screen/breathalyzers	✓		
11.	I made sure the caregiver tested the child for alcohol or drug use by breathalyzer or drug screen			✓

[a] Therapist-reported CM-TAM items. Youth- and caregiver-reported CM-TAM items use “the therapist” in place of “I” and “my child” in place of “the child.”

[b] Study 1 = Juvenile Drug Court Integration, Study 2 = Transportability within an Existing Evidence-Based Practice, Study 3 = Transportability to Mental Health and Substance Abuse Treatment Providers.