Fidelity Study of CII's Delivery of Trauma-Focused Cognitive Behavioral Therapy

A Technical Resource for

Improving Service Delivery for Children Affected by Trauma

An Implementation Study of Children's Institute, Inc.

Rochelle F. Hanson
Jason E. Chapman
Sonja K. Schoenwald
Michael de Arellano

August 2016

mdrc
BUILDING KNOWLEDGE
TO IMPROVE SOCIAL POLICY

This material is based upon work supported by the Social Innovation Fund (SIF), a program of the Corporation for National and Community Service (CNCS). The Edna McConnell Clark Foundation's SIF includes support from CNCS and 15 private co-investors: The Edna McConnell Clark Foundation, The Annie E. Casey Foundation, The Duke Endowment, The William and Flora Hewlett Foundation, The JPB Foundation, George Kaiser Family Foundation, The Kresge Foundation, Open Society Foundations, The Penzance Foundation, The Samberg Family Foundation, The Charles and Lynn Schusterman Family Foundation, The Starr Foundation, Tipping Point Community, The Wallace Foundation, and the Weingart Foundation.

Dissemination of MDRC publications is supported by the following funders that help finance MDRC's public policy outreach and expanding efforts to communicate the results and implications of our work to policymakers, practitioners, and others: The Annie E. Casey Foundation, Charles and Lynn Schusterman Family Foundation, The Edna McConnell Clark Foundation, Ford Foundation, The George Gund Foundation, Daniel and Corinne Goldman, The Harry and Jeanette Weinberg Foundation, Inc., The JBP Foundation, The Joyce Foundation, The Kresge Foundation, Laura and John Arnold Foundation, Sandler Foundation, and The Starr Foundation.

In addition, earnings from the MDRC Endowment help sustain our dissemination efforts. Contributors to the MDRC Endowment include Alcoa Foundation, The Ambrose Monell Foundation, Anheuser-Busch Foundation, Bristol-Myers Squibb Foundation, Charles Stewart Mott Foundation, Ford Foundation, The George Gund Foundation, The Grable Foundation, The Lizabeth and Frank Newman Charitable Foundation, The New York Times Company Foundation, Jan Nicholson, Paul H. O'Neill Charitable Foundation, John S. Reed, Sandler Foundation, and The Stupski Family Fund, as well as other individual contributors.

The findings and conclusions in this report do not necessarily represent the official positions or policies of the funders.

For information about MDRC and copies of our publications, see our website: www.mdrc.org.

CHILDREN'S INSTITUTE, INC. PROGRAM EVALUATION
FIXED PRICE SUBCONTRACT (#017545-002)

Original Project Dates: 2/1/13 - 7/31/14 [Notice of Award – 3/18/13]
MODIFIED CONTRACT [11/25/13]: REVISED PROJECT END DATE: APRIL 30, 2015

Final Report
June 5, 2015

Co-Principal Investigators: Rochelle F. Hanson, Ph.D., & Jason E. Chapman, Ph.D.
Co-Investigators: Sonja K. Schoenwald, Ph.D., & Michael de Arellano, Ph.D.
Research Assistant: Carrie Jackson

Table of Contents

Context and Objectives of the Evaluation

The Children's Institute Inc. (CII) was selected to receive a Social Innovation Fund grant from the Edna McConnell Clark Foundation. As part of the funding, each grantee was required to conduct an evaluation of the effectiveness of their program. The purpose of the current research was to conduct a fidelity study of the implementation at CII of Trauma-Focused Cognitive Behavioral Therapy (TF-CBT; Cohen, Mannarino & Deblinger, 2006). CII is a multi-service organization that operates in central and south Los Angeles County to serve families who have been harmed by violence or neglect. TF-CBT, a short-term evidence-based psychotherapy, is commonly used at CII to treat children with emotional and behavioral difficulties as a result of their exposure to traumatic events. MDRC is conducting an implementation evaluation, of which the TF-CBT fidelity study is one component. A team of investigators at the Medical University of South Carolina (MUSC) with expertise in TF-CBT, treatment fidelity measurement methods, and implementation was contracted to evaluate the fidelity with which TF-CBT was implemented with clients at CII. The specific aims of the contracted evaluation are enumerated below.

Aim 1: Review and develop expertise in the use of a newly developed, existing observational instrument to assess adherence to TF-CBT.
Aim 2: Evaluate provider adherence to TF-CBT in treatment sessions as indexed by the selected observational instrument (including psychometric properties).
Aim 3: Assess the psychometric properties of a therapist self-report TF-CBT session checklist [i.e., the TF-CBT Brief Practice Checklist (BPC) currently in use at CII].
Aim 4: Evaluate the magnitude and direction of association between TF-CBT adherence as indexed by the newly developed observational coding instrument and the TF-CBT BPC.

In addition to these core project aims, the study also examined two Exploratory Aims, identified below. These aims are exploratory because the projected size and distribution of the sample of treatment sessions, clients and therapists would not be sufficient to power a more conclusive statistical evaluation.

Exploratory Aim 1: Test in a preliminary way the adequacy of the observational coding instrument for sessions conducted in Spanish as well as those conducted in English. The feasibility of this analysis was dependent on having a sufficient number of Spanish cases recruited into the study.
Exploratory Aim 2: Conduct a preliminary evaluation of relations among therapist variables and fidelity.

Background and Significance of the Evaluation

Taking Evidence-Based Treatments to Scale: The Importance and Challenges of Fidelity Measurement

In addition to addressing the specific aims, the contracted evaluation is expected to contribute to the knowledge base regarding the effective and efficient measurement of fidelity to evidence-based treatments in community practice contexts (Schoenwald & Garland, 2013; Schoenwald, Garland, Chapman, Frazier, & Southam-Gerow, 2011). This is because it encompasses both evaluation of an observational measurement method to assess fidelity to TF-CBT and of a therapist-reported method to do so; and, examines relations among scores obtained using each method. In psychotherapy research, psychometrically sound observational measurements of fidelity are considered the "gold standard." Although used in well-funded treatment efficacy trials, observational methods are generally too complex, time-consuming and expensive to use in community practice settings. In contrast, therapist self-report methods are feasible to use in practice, but are rarely evaluated psychometrically. Further, when self-report methods are evaluated, they are found to suffer from limitations related to statistical interdependencies in data owing to the fact that a single therapist reports on the multiple cases treated by that therapist and do not produce scores consistent with those obtained using observational coding systems.

Because fidelity and its monitoring contribute to the effectiveness and sustained implementation of evidence-based treatments in community practice settings (Aarons, Hurlburt, & Horwitz, 2011; Fixsen, Naoom, Blase, Friedman, & Wallace, 2005; McCleod, Southam-Gerow, Tully, Rodriguez, & Smith, 2013), the need for research on fidelity measurement methods that are both effective and efficient is receiving increased attention

among federal research funding agencies, public service systems, and payers. Effective fidelity measurement methods are those well suited to a specific purpose or end-use (e.g., as a manipulation check on the experimental condition in a treatment trial; or, to train community-based therapists to a particular criterion; or, to produce scores demonstrably sensitive to change that can be used to provide feedback about fidelity to therapists; or, to demonstrate to consumers and payers the extent to which the former received the specific type of treatment sought, etc.) and characterized by evidence supporting their reliability and validity. Efficient measurement methods are those whose use is feasible in community practice. Few methods used to measure fidelity to evidence-based treatments are both effective and efficient. Research is needed to identify the type of information that can be accurately obtained from observational and self-report methods so that informed decisions can be made about the appropriate uses of information emanating from each method. The results of the current evaluation are expected to inform such decisions with respect to the fidelity of TF-CBT.

Assessing Fidelity to TF-CBT

To date, the most effective trauma treatment for youth is TF-CBT (Cohen, Mannarino, & Deblinger, 2006), which addresses trauma-related symptoms including posttraumatic stress disorder (PTSD), depression, and emotional and moderate behavioral problems. It is a manualized, components-based intervention that utilizes techniques from multiple treatment modalities such as cognitive-behavioral, humanistic, and family therapies. TF-CBT has undergone extensive empirical investigation, including at least 15 randomized clinical trials (e.g., Cohen, Mannarino & Iyengar, 2011; Cohen & Mannarino, 1996; Deblinger, Mannarino, Cohen & Steer, 2006; Deblinger, Mannarino, Cohen, Runyon & Steer, 2011). TF-CBT was labeled as a 'Well-Supported, Efficacious Treatment' by the Office for Victims of Crime Guidelines Project (Saunders et al., 2004), given a scientific rating of 1.0 (the highest) on the California Evidence-based Clearinghouse for Child Welfare, and a 3.7 (out of 4.0) dissemination readiness score by the Substance Abuse and Mental Health Services Administration's National Registry of Evidence-based Programs and Practices (http://www.nrepp.samhsa.gov). The age inclusion criterion for TF-CBT ranges from 3 to 18 years. TF-CBT use is widespread in the child welfare and mental health fields.

A recent study found that 78% of respondents to a survey of clinicians providing mental health services to children had been trained in TF-CBT and reported using it on a regular basis (Allen & Johnson, 2012). However, findings from the same study suggested that many of those providing TF-CBT were not delivering all of the treatment components and therefore likely not implementing TF-CBT with fidelity. These results are not surprising, given published findings regarding: (a) the variability in community practice of fidelity to evidence-based treatments (see, e.g., Boxmeyer, Lochman, Powell, Windell, & Wells, 2008; Dane & Schneider, 1998; Deci, Santos, Hiott, Schoenwald,& Dias, 1995); (b) contextual factors affecting treatment fidelity, including organizational climate and structure (Schoenwald, Chapman, Sheidow, & Carter, 2009); and treatment-specific clinician training, supervision, or consultation practices (see, e.g., Bearman et al, 2013; Caroll & Rounsaville, 2010; Nadeem, Gleacher, Beidas, 2013; Schoenwald, Sheidow, & Chapman, 2009).

Several approaches have been taken to define and assess fidelity to TF-CBT; however, a single instrument has not yet emerged that has been psychometrically evaluated across multiple TF-CBT trials. As such, organizations like CII that are implementing TF-CBT do not have a standardized and psychometrically sound method to determine if therapists are delivering, and clients are receiving, the treatment as intended. In psychotherapy research, there are three components of treatment fidelity: therapist adherence, therapist competence, and treatment differentiation. Reliable and valid instruments to detect each facet of integrity exist for relatively few treatments (Perepletchikova, Treat, & Kazdin, 2007); and, adherence is the most frequently measured among these. Therapist adherence refers to the extent to which treatments as delivered to clients include prescribed components and omit proscribed ones (Yeaton & Sechrest, 1981). Thus, the core task of adherence measurement is to answer the question "*Did the therapy occur as intended*?" (Hogue, Liddle, & Rowe, 1996, p.335). Accordingly, adherence to TF-CBT was the focus of the current evaluation.

Note, although other aspects of treatment such as dosage, duration of treatment sessions, spacing of sessions, and length of a course of treatment (i.e., a treatment episode) are aspects of treatment that reflect its implementation, and may affect outcomes, they are not typically defined or evaluated in terms of fidelity. Nonetheless, descriptive data regarding these aspects of TF-CBT as implemented during the current evaluation are briefly summarized subsequently.

Method

Recruitment and Consent

All therapist and client recruitment and informed consent procedures were conducted by MDRC and CII. In September and October 2013, MDRC held three webinars for CII clinical supervisors and regional directors to train staff on the procedures for obtaining consent from clients for the research study. Subsequently, CII was responsible for obtaining consent to participate in the research from all new TF-CBT clients and their legal guardians. All clients starting TF-CBT at CII during the study period (i.e., consent period was November 2013-August 2014) were eligible to participate. CII offered all therapists providing TF-CBT during the study period the opportunity to participate in the research. Participating therapists facilitated the consent process for their client(s) by coordinating consent to occur with their clinical supervisor. Participation in the TF-CBT Fidelity study was entirely voluntary and independent of any treatment decisions made by TF-CBT therapists or CII staff. Provision of mental health services was provided by CII regardless of a client or family's decision to participate in this study. Information was not available for the number of clients who declined to participate. However, anecdotal reports from CII indicated that only a few clients declined, primarily because they were unwilling to have their treatment sessions recorded.

A total of 70 therapists were offered the opportunity to participate in the study. Of these, 43 therapists provided initial consent, nine of which later withdrew or were not assigned cases, and 27 declined to participate. Reasons therapists declined to participate included:

- Being in an administrative role (*n* = 6)
- Just learning the model (*n* = 4)
- Time constraints (*n* = 3)
- Left CII (*n* = 3)
- Studying for licensure (*n* = 3)
- On leave (*n* = 2)
- Switched EBPs (i.e., no longer providing TF-CBT) (*n* = 2)
- Serving as a Supervisor (*n* = 1)
- Temporary employee (*n* = 1)
- Dropped when number of therapists was reduced (*n* = 1)
- Only provides treatment to young children (ages 0-5) (*n* = 1)

Participants

Therapists. Of the 70 therapists offered the opportunity to participate, 43 provided initial consent, 9 of which did not participate because they later withdrew or were not assigned cases, and 27 declined for the reasons indicated above. This resulted in a final enrolled sample of 34 therapists, the majority of whom were female (*n* = 30) and identified as Latino (*n* = 20; 58.8%). Regarding race, 8 were Caucasian; 4 were Asian; 1 was African American and the remaining 21 identified as "Other." All of the enrolled therapists (*n* = 34) had graduate degrees; the majority (*n* = 28; 82.3%) had masters' degrees and the remaining 6 had doctoral degrees (*n* = 4 Psy.D.; *n* = 2 Ph.D.). Regarding job title at CII, most were Therapists (*n* = 20; 58.8%) or Senior Therapists (*n* = 4); other job titles included Post-Doctoral Psychologists (*n* = 2); Psychologists (*n* = 2); Intern or Practicum students (*n* = 3) and Supervisor (*n* = 1). Participating therapists were employed at CII for an average of 46.05 months (4.26 years) from their hire date to the end of the study (*SD* = 40.22 months). The average length of time between initial hire date and first TF-CBT case was 18.24 months (*SD* = 33.48 months). The average length of time from TF-CBT training to the first TF-CBT case was 1.08 months (*SD* = 14.38). However, 12 of the 34 therapists saw their first TF-CBT case prior to receiving formal TF-CBT training (ranging from 2 to 44 months prior to participation in a TF-CBT training). The length of time from TF-CBT training to the end of the study ranged from 4 – 97 months (*M* = 26.74; *SD* = 21.06). Since hire date, an average of *n* = 36.65 TF-CBT clients were seen (*SD*=18.51). Finally, the number of study clients per therapist ranged from 0 (*n* = 3 therapists) to 9 (*n* = 1 therapist) (*M* = 3.02 per therapist). Of the 34 initially enrolled therapists, 3 withdrew: 2 were students completing a time limited practicum that precluded their ability to remain in the study, and the third withdrew because of medical leave, resulting in a final sample of *n* = 31.

Clients. The sample of clients who consented to participate included 126 children and caregivers. Of these, 32 clients withdrew from the study, leaving 94 participants in the final sample. The reasons for

withdrawal were: dropped of treatment and/or were closed by CII (*n* = 15) (e.g., client chose not to continue treatment, client refused services; case referred to outside provider); did not meet TF-CBT criteria (*n* = 4); did not want to be audio-recorded (*n* = 4); and client moved (*n* = 3). Additional reasons for withdrawal (for one case each) included: active psychosis, assignment of the case to another treatment approach; client detained; and therapist was unaware the client was enrolled in the study. Two cases withdrew for unknown reasons. The majority of children in the final sample (*n* = 94) were girls (*n* = 56; 59.6%); and most were Caucasian (*n* = 82; 87.2%), with 7 African Americans, 3 Asian, 1 Native Hawaiian/Pacific Islander; and 1 unknown; 81 participants identified their ethnicity as Latino.

Data Sources and Collection

All data for the current evaluation were originally provided by CII and sent to MDRC who then shared the data with the MDRC research team. Three types of data were obtained.

Therapist demographic and experience information. CII personnel collected de-identified data from their human resources and billing records, which included demographics, education, participation in TF-CBT training, length of employment and supervision at CII, as well as general and TF-CBT caseloads. Preliminary data were collected in November 2014; final data were collected and shared with MDRC and then MUSC in March 2015.

Therapist-reported TF-CBT Brief Practice Checklist (BPC). Therapists completed the BPC after each session to track the progress of each of their TF-CBT cases. Specifically, after each session, therapists endorsed which of the ten TF-CBT components [(Psychoeducation, Parenting, Relaxation, Affective Expression, Cognitive Coping, Trauma Narrative, Trauma Processing, In Vivo Mastery of Trauma Reminders (hereafter referred to as "In Vivo Exposure"), Conjoint, and Enhancing Safety)] and sub-components (total of 31 items distributed across the 10 core components; see Table 7) they provided and also indicated whether or not a caregiver was present for the session. BPCs were then submitted to the clinical supervisors on a weekly basis and reviewed during the weekly TF-CBT supervision meetings. CII administrative staff collected the BPCs from the TF-CBT supervisors and sent them to CII's Research and Evaluation staff who then electronically entered de-identified data from the paper checklists into Microsoft excel files. Each checklist was entered into a separate file, and files were periodically sent to MDRC where research staff compiled them into a single spreadsheet for transmission to the MUSC research team. A total of 108 checklists were received, which included clients retained in the study (*n* = 94), as well as 14 checklists from clients who withdrew from treatment or from the research study. For clients who withdrew, only data from sessions conducted during the study enrollment period were included. A total of *n* = 1,706 TF-CBT sessions were coded by 34 therapists using the BPC. Data were collected from October 2014 through April 2015 for TF-CBT sessions conducted from November 2014 through February 2015.

TF-CBT treatment session audio-recordings. Participating therapists were instructed to record each session (child, caregiver and conjoint) with an enrolled client using an audiotape recorder. After recording a session, therapists confirmed that the session was successfully recorded by marking the appropriate box on the BPC for that session, and then uploaded the recording to a designated location on a CII network drive. CII's Information Technology staff implemented a system that scanned the appropriate network locations on a daily basis and uploaded any new files to a web-based file storage system called Quickdraw. MUSC research team members and coders were provided access to login to Quickdraw to download audiotape recordings of TF-CBT sessions for coding purposes. Each recording filename indicated the following information about the session: client ID, session number, session type, session date, and language. Audiotape recording began in November 2013 and continued through February 2015.

Observational Measurement Evaluation Procedures

Overview. The research procedures required to achieve the aims of the current evaluation included the following: (1) ensuring accurate specification of the TF-CBT adherence components; (2) review and development of the MUSC investigative team expertise in the use of a newly developed coding system and manual for coding the treatment components; (3) training coders in the use of the coding manual and ensuring execution of the coding plan; and (4) analyses of the data resulting from the audio-recordings of TF-CBT treatment sessions at CII and of the therapist-completed checklist, the BPC, using measurement models grounded in Item Response Theory (IRT). Details regarding the nature, advantages, and application of IRT models to evaluate the performance of the observational and therapist-reported TF-CBT are provided in the Data Analysis Strategy section of this report and throughout the Results section.

TF-CBT Version of the Therapy Process Observational Coding System for Child Psychotherapy (TF-CBT TPOCS-S; Deblinger, Dorsey, Cooper, McLeod & Garland, 2013)

The original evaluation proposal specified ground-up development of an observational coding system to index adherence to TF-CBT. In the months that transpired between submission of the proposal to MDRC and the contract award, however, the MUSC investigative team identified and cultivated an opportunity to utilize an observational coding system for TF-CBT that was being evaluated in a multi-year randomized trial funded by the National Institute of Mental Health. That trial focuses on the effects of clinical supervision strategies on various aspects of the implementation of TF-CBT by a large sample of community-based therapists in Washington State (Dorsey et al., 2013). The study investigators include Dr. Esther Deblinger, a developer of TF-CBT and co-author with several colleagues -- including Dr. Hanson -- of an early version of a TF-CBT coding system (Deblinger, Cohen, Mannarino, Runyon, & Hanson, 2008).

The TF-CBT TPOCS-S (hereinafter TPOCS-S) is designed to index the content of the 10 components of TF-CBT: Psychoeducation/Parenting; Relaxation; Affective Identification and Modulation; Cognitive Coping; Trauma Narrative and Processing; In Vivo Exposure; Conjoint; and Enhancing safety; as well as assessment and other topics/crisis or case management. The acronym for the 10 TF-CBT core components is PRACTICE, and they formed the basis for the afore-mentioned early coding system, the PRACTICE Treatment Adherence Checklist, Original Tape Coding Form (Deblinger et al., 2008). The TPOCS-S also includes items that index a variety of therapeutic strategies derived from multiple theoretical approaches to psychotherapy (e.g., establishing an agenda, use of Socratic questioning, teaching, reflective listening), revised for the Washington study to reflect strategies clinicians might use when providing TF-CBT. The focus of the current evaluation is on the adherence of clinicians to the content of TF-CBT components, and thus on the items in the coding manual that pertain to this content. These items are coded for three types of sessions: caregiver only, child only, and conjoint caregiver-child sessions. After the full session is coded (at five minute intervals), coders provide a rating of "extensiveness" (a six point rating to reflect the thoroughness or intensity of the intervention) for each item coded (by session type). In the ongoing study in Washington, coders also rate the therapeutic alliance and therapeutic technique items. Since therapeutic alliance is not a component of adherence, it was not coded in the current evaluation. Therapeutic technique items were coded; however, because these were crafted for the Washington study and their association with the content of TF-CBT PRACTICE components will be evaluated for the first time in that study, analyses were not conducted of data from these codes. The TPOCS-S manual provides detailed, comprehensive instructions for the coding system. Each item detailed in the manual includes the code, a brief description of the code and its purpose within the scale, exemplars (i.e., examples of therapist statements reflecting the code), frequently co-occurring codes, and guidelines for differentiating the code from other TPOCS-S codes. The manual also provides a description of the recommended steps for coder training; these procedures were utilized in the revised protocol for the current evaluation and are described in the next section of this report.

The use of the TPOCS-S coding system confers upon the current evaluation several benefits, which include the following: (1) leveraging of the significant federal research funds that supported the development of the observational coding system and manual at the University of Washington; (2) leveraging of the expertise of the investigators who designed the coding system (one of whom is a developer of TF-CBT) to support the development of expertise in its use among the MUSC investigative team; (3) demonstration of the feasibility and efficiency of replicating procedures to train coders in the observational coding system; and (4) use of the same observational measurement method to assess adherence to the TF-CBT treatment components in two studies involving distinct samples of children, therapists, and organizations; thereby facilitating future comparison of findings across samples. The findings of this component of the evaluation of CII services are less likely to be idiosyncratic (a common feature of program evaluations) and more likely to contribute to the evidence base regarding the fidelity of TF-CBT in community settings and its measurement.

Coder training procedures. On July 31 and August 1, 2014, six individuals with a master's degree or higher and expertise in TF-CBT participated in a two-day in-person and web-based (via HIPAA-compliant Adobe Connect) training. This was comprised of didactic instruction and discussion of TF-CBT and of the coding manual, group coding of audio-recorded TF-CBT components and audio-recorded full sessions, and individual coding of audio-recorded full sessions. The training was conducted at MUSC, and was facilitated by Dr. Shannon Dorsey, co-author of the TPOCS-S and Principle Investigator of the Washington study in which it is being used, two master coders from that study, and MUSC investigators, Drs. Rochelle Hanson and Michael

de Arellano. Prior to the in-person training, coders completed assigned readings. The procedures for coder training and quality assurance were informed by those used by Drs. Chapman and Schoenwald (1R21MH097000), and by Dr. Dorsey and the University of Washington coding team. After the initial, two-day training, coders participated in a 2-month (August – September) training period that took place in group and individual formats, facilitated by the MUSC investigators and University of Washington TPOCS-S experts. This 2 month training period included individual coding and participation in weekly small group telephone conferences with the study investigators and periodically with the consultants until they reached adequate reliability (ICC = .80) and thereby successfully completed the certification process. The table below details the specific training activities and time that was required for each component.

Table of Coder Training Activities

CODER TRAINING	
Training Activity	**Duration (hours)**
Read Manual	8
In-Person Training	16
Independent Coding	32
Group Coding Review	8
Certification	10
Total	74 (12 hrs/wk for 6 wks)

Following the in-person training session, Wave 1 training involved coding of 6 tapes during consultation calls with the project investigators and University of Washington consultants. Mean *ICC* across coders was .98; during Wave 2 training, 10 tapes were coded independently as "Gold Standard" for certification (benchmarked against the University of Washington master codes) and mean *ICC* = .978, exceeding the *a priori* certification standard of .80. After certification, coders transitioned to independent coding of session tapes with ongoing quality assurance to prevent coder drift.

For the training-related activities, study investigators were responsible for downloading audio-recordings from the CII secure file hosting website directly to HIPAA compliant USB devices purchased specifically for this project. Study investigators identified the session segments that were used in the coder training.

Audio coding procedures. Following certification, coders began independently coding each of the audio recordings. Sessions to be coded were randomly assigned to coders (the coding plan and assignment procedures are described in detail below). Coders downloaded assigned audiofiles from the CII secure file hosting website, Quickdraw, directly to their HIPAA compliant USB devices. The recordings were coded directly from this device, and after being coded, they were deleted from the device. Codes were recorded on paper-and-pencil coding sheets that were developed for the coding system. This form included a field indicating the language of the treatment session (i.e., English or Spanish). Project personnel entered records into a Microsoft Excel database prior to data analysis, and the database was housed on MUSC secured server space.

During the coding phase, the coders consulted with the investigators, and periodically with the training consultants, to prevent coder drift. Reliability coefficients and other coder statistics were calculated on a routine basis to determine the need for booster training. Approximately 15% of the sessions were double-coded for reliability purposes, as is consistent with published research involving observational coding of adherence in psychotherapy studies, in which 10 - 20% of sessions are coded twice for reliability purposes. The time required for coding each session was estimated to be approximately 5-10 minutes longer than the actual session length, resulting in an average of 48 minutes (*SD* = 16.9) per session tape.

Coding plan. Guided by theoretical and statistical estimates, the coding plan was carefully balanced across the number of therapists, the number of clients per therapist (and the total number of clients), and the number of sessions per client. Accordingly, having approximately 30 therapists allows for stable and precise statistical estimates. Likewise, having 4 clients per therapists provides a sufficient sample of therapists' implementation of TF-CBT to provide trustworthy estimates. Finally, because the components of TF-CBT are implemented at different times during the course of treatment, it was important to adequately sample the full course of treatment for each client. This balanced coding plan makes it possible to answer questions about the extent to which the implementation of TF-CBT occurs primarily at the level of the therapist (i.e., each therapist

has a stable, characteristic level of TF-CBT fidelity), at the level of the case (i.e., the therapist's fidelity is largely determined by the case being treated), at the level of the session (i.e., for each therapist and case, the level of fidelity fluctuates from session to session), or at some combination of each of these.

Final sample of recordings. To determine the final sample of recordings, it was first necessary to complete extensive cleaning of the list of uploads to the Quickdraw site. Each uploaded recording was logged by research staff in an Excel database; and for data cleaning and preparation for analysis, this database was converted to SPSS (v22) format. When the final, main wave of data cleaning was completed as of January 28, 2015, the total number of uploads on Quickdraw was 1,145; however, these were not all unique, independent uploads, and to make this determination, the fields from the file name were used. Specifically, the file naming convention included fields for the client ID, session number, session type, session date, and session language. However, 156 of the uploads were missing one or more of these fields, and 206 of the uploads provided some evidence of being a duplicate upload (e.g., duplicate file name, multi-part upload, repeat upload due to previously identified file errors, repeat uploads due to previously identified missing fields). The cleaning process involved a number of steps, including: a manual review of potentially duplicate sessions, estimating language (if missing) based on other uploads from the same client and/or therapist, resolving session and date discrepancies in repeat uploads, and assigning available conjoint sessions in place of parent sessions in cases with fewer than 8 caregiver sessions, etc. After completing these steps, the total number of unique uploaded recordings was 1,009. Of these, 722 were assigned for coding, including 195 Spanish language sessions covering 36 clients, and 105 were assigned for evaluating inter-rater reliability. In March of 2015, an additional 500+ recordings were uploaded to the Quickdraw site, however, due to the imminent deadline for completion of the evaluation, the approximately 200 sessions eligible for assignment were not assigned. The TPOCS-S results reported below are based on 587 coded sessions, of which 80 were sessions coded for inter-rater reliability, and 68 were Spanish sessions. The number of clients coded per therapist ranged from 1 to 5, with an average of 2.9 (*SD* = 1.3). The number of sessions coded per client, including both youth and parent sessions, ranged from 1 to 13, with an average of 7.0 (*SD* = 3.4).

Coding assignments. The coding assignments occurred in two main waves, the first began in late September of 2014 and the second in mid-February of 2015. As noted previously, for each case, a maximum of 8 child and 8 caregiver sessions were assigned and coded, and as such, the total sample of available, unique recordings was not coded. Generally, recordings to be coded were randomly assigned to coders, with two main caveats. First, uploads of sessions conducted in Spanish were randomly assigned to one of two bilingual coders (who also coded sessions conducted in English). Second, to ensure that the coded sessions provided a thorough evaluation of the full course of TF-CBT, the less frequently occurring later sessions were prioritized for random assignment relative to the more common earlier sessions. Also, as mentioned above, ~15% of the assigned recordings were also assigned to a second coder for the purpose of evaluating inter-rater reliability.

Data Analysis Strategy

For evaluating the performance of measurement instruments, models based in Item Response Theory (IRT), referred to as modern test theory, are recommended by the *Standards for Educational and Psychological Testing* (*SEPT*; AERA, APA, & NCME, 2014), Wilson (2005), and Wolfe and Smith (2007). As detailed by Smith (2000) and Bond and Fox (2007), IRT models improve in important ways upon traditional models based in Classical Test Theory. Among these improvements, IRT models provide separate information about both items and respondents. With traditional measurement models, these are confounded. For example, for adherence measurement using traditional methods, if the resulting scores are low, it is not possible to determine whether therapists used few of the intervention components or whether the components being assessed were too advanced. Additionally, the information about items and respondents provided by IRT models is on a common scale, and this permits direct comparisons. Thus, it is possible to straightforwardly estimate the probability of a specific therapist implementing a specific intervention component. Further, adherence measurement often employs dichotomous and/or ordered categorical ratings. With traditional models, the options for modeling such data are limited (e.g., based on the tetrachoric correlation matrix), or the data must be assumed to be continuous, on an interval-scale, prior to analysis. This assumption means that each category is meaningfully distinct and equally spaced (e.g., that the distance between a rating of Not Used and Low Extensiveness is identical to that between Moderate Extensiveness and High Extensiveness). Because IRT models are logit-based, this issue is directly addressed. Rating scales are modeled in their original form, with the distance between pairs being unevenly spaced as indicated, and the models provide a

formal evaluation of the rating scale's performance. Related to this, the scores that result from IRT models are true interval-scale scores; they are not based on the numbers used to represent ordered categories. This means that the "scores" that result from IRT analysis are suitable for mathematical and statistical operations. Finally, IRT models are highly flexible, they readily accommodate missing data and a variety of rating scales within a single instrument, a variety of item types and distributions, and they provide a wealth of output for thoroughly evaluating an instrument's performance. With traditional methods, much of the evaluation is limited to internal consistency, factor structure, and correlations; however, IRT models provide multiple indicators of dimensionality, the fit of items to the model, rating scale performance, reliability and separation, person fit to the model, and so on.

For the aims of the present evaluation, before addressing the implementation at CII of TF-CBT, it is essential to establish that TF-CBT adherence and extensiveness are being measured as intended. According to the *Standards for Educational and Psychological Testing*, this requires far more than a brief work-up of psychometrics; rather, it requires measurement developers and users to build a case for the valid use of the scores based on multiple sources of evidence. The analytic strategy detailed next, and the organization of the results, is intended to provide such evidence, or the lack thereof, for the measurement instruments used to evaluate the implementation of TF-CBT.

A special case of an IRT model is the Rasch measurement model, and this model was used to evaluate Aims 2 and 4 for the TPOCS-S and Aim 3 for the Brief Practice Checklist (BPC). The Rasch model is expressed by $\ln(P_{ni} / [1 - P_{ni}]), = b_n - d_i$ (Rasch, 1960). According to this model, the probability of a TF-CBT component being implemented by a therapist, P_{ni}, is the net result of the therapist's level of adherence, b_n, and the difficulty/applicability of the TF-CBT component, d_i. This means that an expert TF-CBT therapist would have a high probability of implementing a basic component of TF-CBT, for example, Psychoeducation about trauma. Likewise, a novice therapist would have a low probability of implementing an advanced or less common component of TF-CBT, such as In Vivo Exposure. Following from this model, with separate information about the TF-CBT components and the therapists implementing the components, the results provide a substantial amount of information for evaluating the quality of the measurements provided by the instrument.

The main measurement models were specified using WINSTEPS and FACETS software (Linacre, 2014a; Linacre, 2014b), and mixed-effects formulations of the same models (i.e., hierarchical generalized linear measurement models; e.g., Beretvas, 2005; Wang 2007) were specified using multilevel modeling software, primarily HLM (Raudenbush, Bryk, & Congdon, 2013). Following the methods described by Bond and Fox (2007), Schumacker and Smith (2007), and Linacre (2002), and as recommend by the Standards for Educational and Psychological Testing, the models evaluated reliability and validity evidence based on: (a) the degree of dimensionality (i.e., whether an instrument has multiple factors or sub-scales) in the data, (b) the fit of the items and therapists to the model, (c) the performance of the rating scale, (d) the precision and reliability of the item and person scores, (e) the ability of trained raters with TF-CBT expertise to provide consistent, accurate observational ratings, and (f) the correspondence between the empirical and theoretical ordering of item difficulties. Each of these is briefly described in the Results section, including the key considerations related to analysis and interpretation. Supplementary IRT-based item bifactor models implemented in IRTPRO software (Cai, Thissen, du Toit, 2011) were used to evaluate the influence of dimensionality (Gibbons, Bock, Hedeker, et al., 2007; Reise, Morizot, & Hays, 2007). Please note that these models were used to evaluate TF-CBT fidelity data obtained from the observational instrument (TPOCS-S) and from the therapist-reported BPC. In addition, for the both the TPOCS-S and BPC, per the aims of the evaluation, the models were used first to evaluate session-by-session data and, then, to evaluate fidelity over the course of a treatment episode. The results ultimately supported measuring and scoring the TPOCS-S and BPC for the treatment episodes, and as such, these results are more heavily emphasized in the sections that follow.

Precision & statistical power. A key benefit of IRT-based measurement models is that the sample size requirements are much lower than for traditional methods (e.g., confirmatory factor analysis), and the models readily accommodate incomplete data. Because measurement models do not focus on detecting differences between groups, "power" reflects the "precision" of the item and person estimates afforded by the sample size. If the sample is quite small, the estimates may be "noisy," or imprecise; if the sample is large, the estimates can be expected to be "stable" and precise. For the Rasch model, stable item and person estimates can be obtained even with small-to-moderately sized samples (Linacre, 1994). Specifically, the sample size of approximately 30 therapists is sufficient for 95% confidence that model parameters will be stable within ±1 logit

(i.e., log-odds units, the scale of measurement for IRT models), which is considered to be sufficiently stable for low-stakes measurement. Of note, however, this estimate is for single occasion, dichotomously scored items; therefore, the proposed repeated measurements and rating scale data offer even greater precision because they drastically increase the sample size and the available information.

Results

As noted in the Background and Significance section, although aspects of treatment such as duration of an treatment episode, number of sessions, frequency of sessions, and session participants are not indicators of adherence, these aspects are often specified in the treatment protocols used in efficacy trials, if not also in the manuals used to train clinicians in the use of a treatment. These aspects of the implementation of a treatment may affect treatment outcomes. Accordingly descriptive information about TF-CBT treatment episodes at CII is briefly presented next, before the results related to TF-CBT adherence.

Treatment Episode Information

Based on the dates of the 1,706 TF-CBT sessions identified via therapist-reported BPC data, the average duration of TF-CBT treatment episodes was 18.79 weeks (*SD* = 7.66), during which time clients participated in an average of 15.75 treatment sessions (*SD* = 5.08). Of these sessions, approximately 63.4% (*n* = 1,065) were with children only, 35% (*n* = 588) with caregivers only, and 1.6% (*n* = 27) were conjoint sessions. These descriptive data indicate that the frequency and dosage of TF-CBT delivered at CII fall within the parameters of the treatment protocol, in which the frequency and types of sessions are flexibly specified as including 8-20 sessions comprised of individual parallel sessions with the child and caregiver, as well as conjoint sessions (Cohen, Mannarino, & Deblinger, 2006; 2012; Pollio, McLean, Behl, & Deblinger, 2014.) To the extent that child and caregiver sessions are intended to occur in parallel, one might reasonably have expected a more even distribution of child and caregiver sessions than was found in the current sample; and, the number of conjoint sessions appears lower than expected. Indeed, in randomized trials evaluating the effectiveness of TF-CBT, the protocol does specify that sessions should be approximately equally distributed between child and caregiver individual sessions, with at least two conjoint sessions for sharing of the Trauma Narrative and Enhancing Safety. However, as noted in previous research (e.g., Cohen et al., 2011; Deblinger et al., 2011) and two recent reviews (Cary & McMillen, 2012; de Arellano et al., 2014), caregiver participation is a frequent challenge to TF-CBT implementation, which may explain the lower than prescribed frequency for caregiver and conjoint sessions.

Aim 1. Review and develop expertise in the use of a newly developed, existing observational instrument to assess adherence to TF-CBT.

There are no measurement or inferential statistical models associated with Aim 1.

Aim 2. Evaluate provider adherence to TF-CBT in treatment sessions as indexed by the observational instrument (including psychometric properties).

Before evaluating the performance of a measurement instrument, it is essential first to define the instrument's intended use (*SEPT*, 2014). "Intended use" refers to the decisions that will be made based on the instrument's scores, and all reliability and validity evidence must be evaluated with reference to this use. Consequently, "reliability" and "validity" reflect a specific use of an instrument; that is, they do not reflect a stable, inherent property of an instrument. To illustrate this, an oven thermometer may be reliable and valid for measuring oven temperature, but it may not be reliable or valid for measuring freezer temperature. Extended to adherence measurement, a brief series of therapist self-rated questions may provide a score that can be validly used to inform routine supervision of the therapist's implementation of an evidence-based treatment. However, the same questions may not necessarily be used to make decisions about pay-for-performance. For this higher stakes decision, a more extensive, more formal measurement system would likely be required. This raises the question: *What is the planned use of the scores from the TPOCS-S or BPC*? For this evaluation, it is assumed that the planned use is to evaluate the adherence to TF-CBT as implemented by therapists with clients at CII, which could include evaluation of individual sessions or over the course of treatment (i.e., a treatment episode). As such, the results do not speak to the valid use of the TPOCS-S in "hiring or firing" or other high stakes employee decisions. With the end-goal of evaluating adherence to TF-CBT at CII, it is first essential to determine whether the measurement instrument is performing as intended. To do this, a variety of

sources of evidence were considered, each of which is described in the sections that follow. The TPOCS-S items are provided for reference in Table 1.

Inter-Rater Agreement

The first step toward evaluating the TPOCS-S is to determine whether the coder training procedures and coding protocol were effective and implemented as intended. If this is the case, it means that two raters who independently code the same recording should provide the same ratings for each PRACTICE element. Although there are a variety of methods for evaluating inter-rater reliability, in IRT, the primary method is to compute the level of absolute agreement across items rated by two or more raters. Correlational approaches are an alternative, however, their significant limitation is that raters can achieve a high level of reliability even if their ratings are consistently different by one point on the rating scale (i.e., if one coder consistently codes 1 point lower). For fidelity measurement, this is a significant limitation because raters are expected to perform as "coding machines" (Linacre, 2014a). That is, rather than providing an expert interpretation or judgment, fidelity coders are expected to document the occurrence and extensiveness of the components in an identical manner. Ideally, according to Linacre, "rating machines" will agree at a rate of ~90%.

Inter-rater agreement was calculated using a Many-Facet Rasch Model (MFRM). This model included "facets" for the therapist, client, session, raters, and items. The opportunities for agreement were calculated as the cases with matching therapists, clients, session numbers, and item numbers. Of 828 opportunities for agreement, there were 612 (74%) exact agreements. This level of agreement is acceptable, though somewhat low. However, as detailed subsequently in the Rating Scale Functioning section, the rating scale did not perform as intended. This means that some categories were not meaningfully distinct from adjacent categories. After adjusting the rating scale to address these issues, the level of absolute agreement increased to 697 of 828 (84%). An additional indicator of the quality of rater performance is provided by the reliability of rater facet in the model. Although counter-intuitive, with the MFRM, the rater facet should have *low* reliability, as this indicates that it is not possible to differentiate the raters based on their rating styles. In this case, coder reliability was low at .34, which reflects a consistent, shared style across raters. *In sum, these results provide evidence that the coder training procedures and coding protocol were implemented as intended and provide confidence in the results that follow*.

Session-by-Session Results

The TPOCS-S yields codes for TF-CBT components that are implemented in a single session. Because the measurement occurs at the level of sessions, it is important first to evaluate the performance of the instrument at that level. Based on these results, a decision must be made about the appropriate level for ultimately scoring, reporting, and using the scores. This is especially true for fidelity measurement. For example, if the planned use of an instrument is to provide ongoing feedback to a therapist so as to improve the delivery of the intervention with a specific client, validity evidence for this use must occur at the level of individual sessions. Likewise, evaluation of the performance of the instrument may indicate that most of the information about fidelity has to do with clients; that is, that fidelity reflects a course of treatment for a client but varies considerably from client to client within a given therapist. However, the information about fidelity could also reflect a stable, characteristic level of implementing an intervention for individual therapists. If this is the case, it would suggest that therapists do not vary in how they use the intervention. For quality improvement efforts, it is essential to determine which of these is most likely to be the case, as efforts to improve the delivery of the intervention will vary depending on the answer. Finally, when changing data from their original form, for example, taking session data and aggregating to a therapist-level format, assumptions and decisions are imposed that may or may not be justified. Therefore, as much as possible, these decisions should be guided by empirical evidence for the instrument performing at that level.

Across a wide variety of analyses performed, there was consistent evidence that the TPOCS-S does not provide strong measurements of TF-CBT fidelity at the level of individual sessions, with the most significant problems being low reliability and poor targeting of items to individual sessions (i.e., most items were very unlikely to occur in a given session). The descriptive percentages of sessions with each component endorsed are reported in Table 2. To inform the decision about the appropriate level of aggregation of the data, analyses were performed to estimate the proportion of variance in session-level TPOCS-S ratings that was attributable to components (i.e., items), sessions, clients, and therapists. The model, however, did not converge, reflecting that nearly all of the variance was attributable to the items. Based on this finding, three alternative data configurations were considered: (1) modeling based on theoretical *treatment phases* (i.e., with PRAC

components in sessions 1-6, TI components in sessions 7-12, and CE components in sessions 13+), (2) modeling in a *cumulative* manner from session to session, where a PRACTICE component, if previously implemented, receives credit in future sessions, or (3) modeling across the full *course* of treatment, with the components indicating whether a PRACTICE component was ever implemented. The results from the model evaluating the *course* of treatment were most promising and are reported here in detail. The descriptive percentages of clients receiving each component, along with the average extensiveness rating on the revised 3-point scale (ranging from 0-2; see Rating Scale Functioning) are reported in Table 3. Of note, the models evaluating measurement by phase did not perform well, whereas the model evaluating cumulative scoring was viable but not as strongly performing as the model for the course of treatment. To configure the data for the course of treatment, for a given therapist and client, across all sessions coded, the maximum observed rating for each component was retained. An alternative approach would be to use the mean; however, the maximum rating was thought to better reflect the extensiveness of implementation of that component with the respective client. The resulting data structure was such that each client had a single row of data.

Course of Treatment Results

Dimensionality. With measurement models based in IRT, the first step is to evaluate dimensionality in the data. This is like performing a factor analysis using methods from Classical Test Theory, and with either approach, the goal is to determine whether an instrument has multiple factors, sub-scales, etc. If there is substantial dimensionality, then a different extensiveness score must be estimated and reported for each of the dimensions. Dimensionality is not necessarily good or bad, but if it is present and ignored, it can produce misleading findings. An example of this would be an academic achievement test that provided a single score rather than separate scores for reading and math. This is a problem because it is possible to have a high score on one and a low score on the other. For a behavioral substance abuse treatment, there could be separate dimensions for the cognitive-behavioral treatment components and the components related to drug testing and the use of consequences. It would be possible for a therapist to have a high score on one but not on the other, thus reflecting the importance of scoring and reporting the dimensions separately. With the TPOCS-S, evaluating dimensionality is challenging for a number of reasons: (1) the items are rated on a seven-point ordered categorical rating scale; (2) all items are not applicable in all sessions, but specific items are not rigidly linked to specific sessions; (3) there is a general sequencing of components, but the sequence is not rigidly prescribed; (4) the items cover a number of distinct content areas, but each of these has only one item; (5) some items may occur one time and not again, whereas others may occur repeatedly; (6) the items are completed longitudinally, on a session-by-session basis; and (7) there are only 12 total items.

The primary method of evaluating dimensionality was a principal components analysis (PCA) of standardized Rasch residuals from a Rasch rating scale model; however, supplementary analyses were performed that included exploratory bifactor models (in Mplus software), and confirmatory bifactor models (in IRTPRO software). The PCA of standardized Rasch residuals is substantively different than traditional PCA. In the Rasch model, there are parameters (i.e., scores, statistics) for both items and persons, and for the TPOCS-S, the “persons” are individual sessions. A certain amount of the variance in the TPOCS-S ratings can be explained by the model, and the remainder is residual variance. The Rasch PCA is performed on this residual variance. If there is no dimensionality, the residual variance should be random. Therefore, to assess dimensionality, the PCA attempts to find structure in the residual variance. There is always some level of dimensionality in data, and as such, the key question is whether there is enough to warrant separate analyses for each dimension. To determine which items form which dimensions, the model specifies a series of contrasts in the standardized residuals, and each contrast separates the items into two groups—those loading positively with the contrast, and those loading negatively with the contrast. The direction of this loading is inconsequential, but for the first contrast, the items that group together on each side (i.e., positive or negative) are suggestive of the first two dimensions.

There are three main indicators of the level of dimensionality. The first is the total variance explained by the item and person scores. Ideally, if there is no meaningful dimensionality, this value will be ≥ 60%, and for the TPOCS-S, it was 59.1%. The next is the eigenvalue of the first contrast. Based on conventional standards, if there is no meaningful dimensionality, this value should be ≤ 2.0, and for the TPOCS-S, it was 2.5. The third is the percentage of unexplained variance that is attributable to the first contrast. If there is no meaningful dimensionality, this value should be ≤ 5.0%, and for the TPOCS-S, it was 8.6%. The results generally indicate that there is borderline meaningful dimensionality. The percentage of explained variance is excellent, but the eigenvalue and the percentage of unexplained variance due to the contrast are both slightly above the rules of

thumb. As displayed in Table 4, the first suggested dimension is formed by components Psychoeducation, Relaxation, Affective Expression, and Assessment. The second suggested dimension is formed by Trauma Narrative, Trauma Processing, Conjoint, and In Vivo. The loadings for Other Topics, Parenting, Cognitive Coping, and Enhancing Safety were inconclusive. The correlation between the client scores based on the two separate groups of items is .61, indicating that largely the same, though not identical, conclusion would be reached based on the two groups of components. To further dimensionality, confirmatory and exploratory item bifactor analyses were performed. With the item bifactor model, every component loads on one general dimension and also on one or more smaller, specific dimensions. If there are distinct dimensions, the loadings on the general dimension will be low, and the loadings on the specific dimensions will be high. If there is a single, general dimension, the opposite will be true. In the confirmatory model, the results indicated that Trauma Narrative, Trauma Processing, In Vivo Exposure, and Conjoint may form a specific factor. However, the exploratory bifactor results conflict with this. In the exploratory model, all items, with the exception of Other Topics, load strongly on the general dimension.

In making a final decision about dimensionality, a critical consideration with the TPOCS-S is that there are only 12 items. Therefore, separating the components into two dimensions, while addressing potentially meaningful dimensionality, would have the cost of degrading measurements in other ways. For example, based on the components forming the first possible dimension, reliability drops from .77 to .55. For the second suspected dimension, reliability is maintained, but the components target the sample very poorly. “Targeting” is thoroughly described in a subsequent section, but in this case, the small group of components forming the second possible dimension is only sufficient for assessing clients who experienced either very high or very low levels of TF-CBT fidelity, but there are no components targeted to clients receiving average levels of fidelity. Given the small number of components, and the measurement limitations that result from sub-dividing them, the decision was made to proceed with measuring and scoring the TPOCS-S as a single dimension of TF-CBT fidelity.

Item Fit. After determining that the TPOCS-S may be reasonably treated as a single score, the next step is to determine whether there are components that serve to degrade the measurements. The model provides several types of item fit statistics that can be used to identify such components, those providing more “noise” than precision to the extensiveness score. For the ultimate goal of evaluating TF-CBT fidelity, this is an essential piece of validity evidence. Of four possible fit statistics, the standardized Outfit statistic has the strongest evidence supporting its use (Smith, 2000). This statistic is sensitive to unexpected ratings on a particular item. An example of this type of misfit would be a spelling test with a low score, in which the student spelled the most difficult word correctly. For TF-CBT, each client has an extensiveness score. If a client experienced low extensiveness, then an unexpected rating would be high extensiveness on one of the most advanced and rarely occurring components. The reverse is also true, if a client experienced high extensiveness, then all of the easier and more commonly occurring components should have been delivered during the course of treatment. Outfit detects items with this tendency. Values in excess of +2.0 are considered significantly misfitting. For misfitting items, before using the instrument in the future, one of several steps may be taken: (1) delete the item from the analysis and revise the item for future use around the suspected source of misfit, (2) model the item on a different dimension, (3) delete the item from the analysis and from future use, or (4) retain the item as-is due to the importance of its content. The standardized Outfit statistics for the TPOCS-S are reported in Table 5. One component was identified as significantly misfitting: (C.03E) Psychoeducation. Psychoeducation is one of the more commonly occurring components, and as such, the misfit suggests that there is a tendency for courses of treatment with high extensiveness not to have an indication of Psychoeducation occurring. There are several possible explanations for this. It could reflect the sampling plan (i.e., due to random assignment of sessions, sessions in which Psychoeducation occurred were not sampled thoroughly); it could reflect some therapists implementing components with high fidelity but without explicit use of Psychoeducation; or it could reflect Psychoeducation being an aspect of several other TF-CBT components (i.e., despite Psychoeducation occurring, it was subsumed under another TF-CBT component). Of note, however, the misfit statistic was 2.1, only slightly above the threshold; likewise, two of these explanations reflect features of the coding system rather than CII’s implementation of TF-CBT. As a result, the decision was made to retain this component in the models.

Rating Scale Functioning. The next step is to determine whether the rating scale performs well or needs adjustment prior to producing extensiveness scores. The TPOCS-S is rated on a 7-point ordered categorical rating scale. The rating scale construct, extensiveness, reflects the thoroughness and frequency of

the PRACTICE component being delivered. For the 7-point scale, the category values and labels are 0 (Not Used), 1 (Low), 2 (Low), 3 (Medium), 4 (Medium), 5 (High), and 6 (High). For the ultimate goal of evaluating TF-CBT fidelity, the rating scale provides an important source of validity evidence for the planned use of the fidelity score. Specifically, with traditional measurement models, the rating categories' values are treated as true interval-scale numbers that can be added, averaged, etc. However, this assumes even spacing between categories (that the distance *Not Used* to *Low* is the same as the distance from *High* with a rating of 5 and *High* with a rating of 6), and it does not address whether there is evidence to support 7 meaningfully distinct categories that were understood and utilized as intended. Measurement models based in IRT treat the ratings as ordered categories, allowing for varying distances between adjacent categories, and because these models are "logistic," they result in true interval scale scores that are appropriate for mathematical and statistical operations. A formal rating scale analysis is performed to determine the extent to which the rating scale performed as intended, as evidenced by indicators that raters have a shared understanding of the categories, utilize the categories in a consistent manner, and that each category is meaningfully distinct from the other categories. The rating scale analysis followed the recommendations of Wright and Master (1982) and Linacre (2002). The results indicated that 32% of the ratings were 0 (i.e., Not Used), 8% were 1, 10% were 2, 11% were 3, 14% were 4, 19% were 5, and 6% were 6. The results also indicated that categories 1 to 4 were not meaningfully distinct.

The performance of the rating scale is illustrated by the category probability curves in Figure 1. The y-axis is the probability of a response in a particular category. The x-axis is more complex. The left side of the x-axis reflects a session with very low extensiveness and a PRACTICE component that is very difficult to do or rarely occurring. As such, it is expected that the component would not be implemented, and reflecting this is the very high probability of a rating of 0 (i.e., the curve on the left for category 0). However, as the level of fidelity in a session becomes more balanced with the difficulty of a PRACTICE element, the PRACTICE element is more likely to occur. As this happens, the probability of a 0 rating decreases, and the probability of a rating greater than 0 increases. For a rating scale that performs well, the probability of a rating of 1 should increase, rise above the curve for category 0, and form its own peak. However, this did not occur. Rather, the probability of a rating of 5 starts to increase and form a peak. This indicates that for a course of treatment with a 50/50 chance of a component being implemented, the rating assigned is most likely to be a 5 (i.e., the next-to-highest rating). On the x-axis, as this balance continues to shift to a session with a high level of fidelity and a component that is easy to implement, the probability of a rating of 5 decreases and the probability of a rating of 6 increases. In this scenario, the component would likely be implemented, and this is reflected by the probability of a rating of 6 being very high. Taken together, these results indicate that the rating scale did not perform as intended. To remedy this, categories 1, 2, 3, and 4 were combined, and categories 5 and 6 were combined, yielding a 3-point scale with values of 0, 1, 2, reflecting "Not Used," "Medium Extensiveness," and "High Extensiveness." The performance of this rating scale is illustrated in the same Figure. In this case, the scale performs well, with 32% of the ratings in the "Not Used" category, 43% in the "Medium Extensiveness" category, and 25% in the "High Extensiveness" category. Each category is distinct and was utilized in a consistent manner across raters, and at some point along the x-axis, each category is the most probable response. When a course of TF-CBT has low extensiveness and the component is difficult, the probability of a rating of 0 is greatest; when the course and the component are matched, the probability of a rating of 1 is greatest; and when the course has high fidelity and the component is easy, the probability of a rating of 2 is greatest.

Reliability & Separation. For the goal of evaluating TF-CBT fidelity, reliability provides evidence for the precision and stability of the estimated fidelity scores for the course of treatment. As detailed by Schumacker & Smith (2007), the Rasch model provides two types of reliability statistics, Rasch reliability and separation reliability. The first, Rasch reliability, ranges from 0.00 to 1.00 and has a similar interpretation as traditional estimates of internal consistency (i.e., the degree to which items in the scale are measuring a similar construct). The second, separation, is arguably more meaningful but is more challenging to describe. Separation provides an indication of reliability because it reflects the capacity of the TPOCS-S to measure meaningfully distinct levels of fidelity. If an instrument is poor, there may be variability in the scores it produces, but the scores may be noisy and imprecise, such that there is little confidence that the score for a given course of treatment is actually different from the score for another course of treatment. Likewise, if an instrument is strong, the scores will be highly precise, which in this case would reflect that the components can be used to discriminate many different levels of fidelity, from the lowest to the highest. The separation value reflects the

number of distinctions that can be made in extensiveness. For the TPOCS-S, separation reliability = 1.9, which means the TPOCS-S can make two distinctions in a continuum of fidelity scores, which is sufficient for meaningfully discriminating 3 levels of fidelity. This finding is consistent with other fidelity instruments that perform well. With the small number of items and the 3-point rating scale, it could have been possible to have a lower level of separation reliability. However, what this reflects is that the items are sufficient for assessing the full range of fidelity, the lowest to highest levels.

Item-Person Map. A key piece of output from the Rasch measurement model is the "item-person map," which is displayed in Figure 2. The map contains a significant amount of information, and although challenging to interpret at first glance, it provides an important summary of the implementation of TF-CBT for the course of treatment at CII. The vertically-oriented numbers on the left are the units for the scale of measurement which, because this is a probabilistic model, are log-odds units (i.e., "logits"). Next to that, also oriented vertically, is the distribution of clients. In this case, each "X" symbol represents a client and the location represents the extensiveness of the overall course of TF-CBT (i.e., the "score" for each client's course of TF-CBT). Clients located toward the top of the distribution received the highest levels of TF-CBT extensiveness (i.e., most items endorsed with the highest extensiveness), and the clients at the bottom experienced the lowest levels of extensiveness (i.e., fewest items endorsed with the lowest extensiveness). On the left side of the vertical dividing line, there is a series of letters: T, S, M, S, T. The "M" is the location of a client with an average course of TF-CBT. In this case, the average is equivalent to an extensiveness score of 1.0, on the 0-2 extensiveness scale, across the PRACTICE components implemented. Above that, the "S" is the location of a session with extensiveness 1 standard deviation (*SD*) above the mean, an average extensiveness score of 1.3 for the components implemented. Likewise, the "T" is the location of a session with extensiveness 2 *SD*s above the mean (an average extensiveness score of 1.6 for the components implemented). The same applies below the mean (i.e., "M"), with "S" and "T" being clients experiencing extensiveness 1 and 2 *SD*s below the mean (average scores of 0.6 and 0.3 for the components implemented). With good measurement, ideally, the distribution of clients will cover a span of ~4 logits, and in this case, the distribution is excellent.

On the right side of the vertical dividing line is the distribution of items, and the interpretation is very similar. The components located at the top are the most difficult or the least strongly endorsed, and the components at the bottom are the easiest, or most strongly endorsed. On the right side of the vertical dividing line, there is the same series of letters. "M" is the location of a component with an average difficulty (e.g., 07E, Cognitive Coping), "S" is the location of a component that is 1 *SD* more difficult than the mean (e.g., 01E, Assessment), and "T" is the location of a component 2 *SD*s more difficult than the mean (e.g., 10E, In Vivo Exposure). The same holds below the mean, and "S" is 1 *SD* below the mean (e.g., 06E, Affective Expression and Modulation Skills). The ordering of these components, if the instrument is performing as intended, should match the theoretical understanding of the construct. One possible explanation for the finding regarding the difficulty of the In Vivo component is that it is only completed if the child is experiencing trauma-related triggers in her/his every day environment (e.g., at home, the neighborhood, in school, in other locations in the community the child frequents). If the clients in the current sample were not experiencing these trauma-related triggers, then it is expected for this component to occur with low frequency. Another possible explanation is that components occurring with low frequency may be more challenging for therapists to execute. Additionally, these findings are consistent with previous research indicating that certain trauma-specific treatment techniques, such as In Vivo, are infrequently used amongst community mental health therapists (e.g., Jensen-Doss, Cusack & de Arellano, 2008).

Finally, a critical feature of the two distributions is that they should be well-aligned with each other, largely overlapping. Thus, the location of the average component should be in about the same position as the average client, which, in this case, is true. Likewise, there should be components that cover the full range of the client distribution. This is called "targeting," and it indicates that the components are well-suited to the sample. As an example, if a math test is "well-targeted" to a class, it is matched to their math skills – few students will miss all items, and few students will get all items correct. For evaluating TF-CBT fidelity in a community practice setting, this is particularly important because therapists may vary in terms of skills, experience, and fidelity. If therapists were using very little TF-CBT, there would be a large number of Xs at the bottom of the figure. Likewise, if most therapists used most components and with high extensiveness, there would be a large group of clients at the top of the map. The orientation of the two distributions indicates that TPOCS-S components are well-targeted to the therapists and clients at CII.

Based on this figure and supplementary model output, it is possible to determine the probability of specific clients receiving specific components with a specific level of extensiveness. The probabilities for a client with an average level of extensiveness are reported in Table 6. Of note, 73% of the sample falls within ±1 *SD* of the mean level of extensiveness, and for clients with higher levels of extensiveness, the probabilities would be higher, and for clients with lower levels, the probabilities would be lower. Ordered from the lowest to highest probability, for an average client, the probability of receiving *some level* of each component was 5% for Parenting Skills, 6% for In Vivo Exposure, 16% for Assessment, 19% for Trauma Processing, 34% for Enhancing Safety, 35% for Conjoint – Prepared Parent/Child for Conjoint Session, 58% for Cognitive Coping, 74% for Relaxation, 80% for Affective Expression and Modulation Skills, 84% for Psychoeducation about Trauma/TF-CBT, 86% for Other Topics/Crisis or Case Management, and 86% for Trauma Narrative. For receiving these components with an *average level of extensiveness*, the probabilities decrease: 4% for Parenting Skills, 5% for In Vivo Exposure, 14% for Assessment, 16% for Trauma Processing, 30% for Enhancing Safety, 31% for Conjoint – Prepared Parent/Child for Conjoint Session, 53% for Cognitive Coping, 70% for Relaxation, 77% for Affective Expression and Modulation Skills, 81% for Psychoeducation about Trauma/TF-CBT, 84% for Other Topics/Crisis or Case Management, and 84% for Trauma Narrative. Likewise, for receiving these components with a *high level of extensiveness*, the probabilities decrease further: 1% for Parenting Skills, 1% for In Vivo Exposure, 3% for Assessment, 4% for Trauma Processing, 8% for Enhancing Safety, 9% for Conjoint – Prepared Parent/Child for Conjoint Session, 20% for Cognitive Coping, 34% for Relaxation, 43% for Affective Expression and Modulation Skills, 48% for Psychoeducation about Trauma/TF-CBT, 52% for Other Topics/Crisis or Case Management, and 52% for Trauma Narrative.

Hierarchical Generalized Linear Measurement Model. This model adds to the previous results by specifically addressing the nested data structure and determining the proportion of the total variance that is attributable to items, clients, and therapists. This information is not provided by the prior models; likewise, the results of this model are largely limited to just this information. As a supplementary analysis, the goal of the model is to determine the level at which the extensiveness of TF-CBT implementation operates. That is, does extensiveness vary from client to client within therapists, or is it largely a stable therapist-level characteristic? Configured for the course of TF-CBT, the data have a three-level structure with items (level-1) nested within clients (level-2) who are nested within therapists (level-3). An indicator for each item was entered at item-level of the model, and no predictors were entered at client or therapist level. The item responses were modeled according to an ordinal outcome distribution with a logit link function. With a random effect for the reference item only, this formulation is a Rasch-equivalent measurement model. Of the total variance in TPOCS-S ratings on the 3-point scale for the course of TF-CBT, 71% was attributable to the items, 29% was attributable to clients, and <1% was attributable to therapists. These results can inform scoring, aggregation, and modeling decisions, and they suggest that the fidelity of TF-CBT as measured by the TPOCS-S can be thought of as a client-level construct. The reliability estimates tell a similar story. Client-level reliability is relatively high, .78. Therapist-level reliability, on the other hand, is very low, .03. In this context, the high value for client reliability indicates that the item responses for a given client are related, such that knowing the rating on an item provides an indication of the overall client score. The low therapist reliability, however, indicates that the scores for clients treated by the same therapist are somewhat independent; knowing the score for one client does not provide much information about the therapist's overall score. Indeed, these results indicate TPOCS-S scores do not index a property of a particular therapist, such as the therapist's overall or average level of fidelity to TF-CBT.

Aim 3. Assess the psychometric properties of a therapist self-report TF-CBT session checklist (i.e., the TF-CBT Brief Practice Checklist) currently in use at CII.

As described for the TPOCS-S, the planned use of BPC scores must be considered prior to analysis and interpretation. For this evaluation, it is assumed that the planned use is as a highly efficient means of documenting TF-CBT fidelity on a session-by-session basis and may also include use as a summary of overall TF-CBT fidelity with a given client or for a given therapist. As with the TPOCS-S, the results presented below are not sufficient for use in high stakes decisions. Ultimately, for evaluating the BPC, the most critical piece of validity evidence will be its correlation with the TPOCS-S. This will determine the extent to which the BPC leads to the same conclusion about TF-CBT implementation as a gold standard, observational coding system. If it does, this is strong evidence for its valid use in measuring TF-CBT fidelity. The BPC items are provided for reference in Table 7, and the descriptive percentage of sessions with each item endorsed is reported in Table

8. When modeled on a session-by-session basis, the performance of the BPC largely mirrored that for the TPOCS-S. As such, these results are omitted, and the results of the measurement models for the course of treatment are presented next.

Course of Treatment Results

Dimensionality. When modeled using data for the course of TF-CBT, the dimensionality results indicate that the item and person measures (i.e., scores) explain 45.4% of the variance, the eigenvalue for the first contrast is 3.6, and the unexplained variance attributable to the first contrast is 6.4%. Likewise, the loadings of the components on the first contrast of the PCA of Rasch residuals are reported in Table 9. The eigenvalue for the first contrast suggests that there is non-trivial dimensionality in the BPC. Given this, the usual procedure would be to separate the items along these dimensions and report the subsequent results separately for each. However, this approach was not taken for the report for two reasons. First, this result could reflect something about the way therapists -- who are not trained in the rating protocol (although they are trained in the treatment) -- rate the items. As such, forming dimensions based on these patterns could be misleading. Second, because the TPOCS-S is a gold-standard observational coding system and the BPC is intended to measure the same construct, the dimensionality results for the TPOCS-S were applied to the BPC. Namely, all items were retained in a single dimension. Because it is intended to measure the same construct, ultimately, the strongest test of the BPC's performance is its correlation with the TPOCS-S.

Item fit. The standardized Outfit statistics for the BPC are reported in Table 10. Six items were identified as significantly misfitting: Therapist provided psycho-education (P1.1), Made normalizing and validating statements (P1.2), Reviewed limits of confidentiality (P1.3), Engaged family (P1.5), Therapist provided parenting skills (P2.1), and Therapist assisted the child in accurately identifying their feelings, and various ways of regulating their emotions (A.1). One source of misfit is likely to be dimensionality in the data. For some items, the second suspected source of misfit is the multi-part item content given the use of "and" or "or" in the items, the inclusion of "e.g.," in the items, and qualifiers such as "if needed." Particularly when relying on untrained raters and individuals providing self-reports, as much as is feasible, it is best to ensure that each item addresses only a single topic. However, as stated, dimensionality is likely a primary source of misfit, and the review of item content does not reveal serious concerns about the structure and wording of the BPC items.

Rating Scale Functioning. Because the BPC data are dichotomous (i.e., no/yes), there was not a formal evaluation of the rating scale.

Reliability & Separation. The reliability estimates for the BPC are strong, with reliability = .82 and separation reliability = 2.14. The BPC items are sufficient for making 2 distinctions in the data, differentiating 3 levels of TF-CBT fidelity.

Inter-rater Agreement. There are no inter-rater agreement data for the BPC because therapists provide self-reports. For the preliminary report, an unanswered question was whether or not therapists can provide accurate self-reports of their implementation of TF-CBT. This question is largely addressed by the evaluation of Aim 4, in which associations between the BPC and the TPOCS-S scores are compared.

Item-Person Map. The item-person map is displayed in Figure 3. In this case, each "#" symbol represents 2 clients and each "." symbol represents 1 client. The clients ("#" and "." symbols) located at the top have the highest levels of fidelity, and those at the bottom have the lowest levels of fidelity. An average client had approximately 28 of the 31 PRACTICE items implemented, with clients 1 and 2 *SD*s above the mean having all items implemented, and clients 1 and 2 *SD*s below the mean having ~20 and ~10 items implemented. The distribution of clients is relatively wide, covering a span of more than 4 logits. The item distribution also covers a wide range. An average item was A.5 (Understanding Intensity of Certain Feelings), an item 1 *SD* more difficult was T.5 (Worked to Modify Cognitive Distortions Throughout the Narrative), and items 1 and 2 *SD*s below the mean were A.2 (Name Variety of Feelings) and R.1 (Therapist Explained the Physiology of Relaxation and Instructed on Methods of Relaxation). For the course-level data, the ordering of items largely reflects the ordering of the PRACTICE components, suggesting that the BPC, scored for the course of treatment, begins with easier elements and progresses to more challenging elements. Regarding the targeting of the items to the sample of clients, the orientation of the two distributions is the opposite of that for the session-level data. Specifically, evaluated across the course of treatment, the elements of the BPC are relatively easy to endorse as having occurred. For example, for an average client, there was a 50% chance of having experienced, per the therapist's report, the least commonly endorsed elements, T.8 (Read the Trauma Narrative to a Caregiver/Supportive Adult) and I.1 (Therapist Developed an In-Vivo Desensitization Plan to

Resolve Avoidant Behaviors). The probabilities for each component being experienced by an average client during the course of TF-CBT are reported in Table 11. If therapists are able to provide accurate self-reports, this reflects a high level of implementation of TF-CBT PRACTICE elements. One measurement limitation, however, is that the clients receiving the highest level of TF-CBT implementation are not targeted by any items, that is, there are no items at their level (see Figure 3). This would be akin to a large number of students receiving a perfect score on a test, and this situation typically reflects the need for additional items that are specifically targeted to the highest performers. Further, compared to the distribution for the gold standard TPOCS-S, based on therapist reports, the BPC client distribution is skewed toward high levels of TF-CBT implementation.

Hierarchical Generalized Linear Measurement Model. The HGLMM results indicate that client reliability = .82 and therapist reliability = .36, with 53% of the total variance attributed to items, 39% to clients, and 9% to therapists. The higher therapist-level reliability estimates and the higher percentage of variance attributable to therapists likely reflect methodological variance. Specifically, when the therapist provides ratings, rather an external, objective, and trained rater, all of the information is filtered through that therapist's perspective. Therefore, there tends to be a level of consistency in that therapist's reports, and this is reflected in the percentage of variance at therapist level. For an external rater of the same course of treatment, the therapist-level ICC likely would be much lower, as was observed for the TPOCS-S.

Aim 4. Evaluate the magnitude and direction of association between TF-CBT adherence as indexed by the newly developed observational coding instrument and the TF-CBT Brief Practice Checklist.

This aim was evaluated in several ways. The session-by-session data for both the TPOCS-S and BPC were used to produce Rasch logit-based session scores, and the correlation between the two scores was computed and graphed. Although session-by-session measurement models were not supported, for evaluating the extent to which the BPC yields the same conclusions as the TPOCS-S, this is an essential first step. With the session by session data, no simplifying assumptions or data aggregation has been imposed. This approach ignores the nested data structure; however, there were two main reasons for this. First, ignoring nesting primarily has the effect of inflating the level of statistical power; however, the focus of this analysis is simply on the magnitude of the correlation, not statistical significance. Second, it is important to evaluate the extent to which the BPC in its original form (i.e., therapist-report on a single session) leads to the same conclusion as the TPOCS-S in its original form (i.e., ratings from a trained coder on a single session). If the two methods of measuring the implementation of TF-CBT lead to the same conclusion, a scatterplot of the pairs of scores should form a straight line, reflecting a high correlation. To the extent that there are departures from a straight line, and a lower correlation, the two methods do not lead to the same conclusion. Of the 443 available coded sessions on the TPOCS-S (English and non-IRR sessions), 424 (96%) matched with available BPC records, and the correlation was small, r (424) = .155. Illustrated by the scatterplot (Figure 4), the two scores do not form a straight line. An obvious concern with therapist self-reports of treatment implementation is that they may inflate, whether intentionally or not, their level of utilization of the various components. The scatterplot speaks to the nature of the disagreements between the BPC and TPOCS-S. Interestingly, for this sample, it appears that there was a tendency for scores from the TPOCS-S to be higher than scores from the BPC, as evidenced by more departures occurring below, and further from, the identity line. This suggests that for single sessions, the TPOCS-S generally reflected higher levels of implementation of TF-CBT relative to the BPC, which is likely attributable to the 3-point, versus dichotomous, rating scale for the TPOCS-S. Because the scatterplot and correlation ignore the nesting of sessions within clients within therapists, to further evaluate the association between the two measurement methods, the unstandardized correlation between the TPOCS-S and the BPC was estimated using a mixed-effects regression model in HLM software. Specifically, TPOCS-S fidelity session scores (level-1) were nested within clients (level-2) who were nested within therapists (level-3), and for the same session, BPC scores were entered as a session-level predictor. The results indicated that there was no significant variability in the association between BPC and TPOCS-S scores from client to client or from therapist to therapist; however, the association between the two scores was statistically significant, $\gamma_{100} = 0.10$, $SE = 0.04$, $T(334) = 2.70$, $p = .007$, but, the magnitude of this association is small, reflecting limited agreement between the two methods when measuring TF-CBT implementation on a session-by-session basis.

Next, the same approach was used, but with scores based on the full course of treatment. The correlation and scatterplot presented in Figure 5 evaluates the extent to which the BPC and TPOCS-S lead to the same conclusion about the level of TF-CBT fidelity received by a client during the full course of TF-CBT. Of

note, the BPC data are far more complete for estimating the overall level of fidelity because the TPOCS-S provides an estimate based on a sample of sessions for each client. As illustrated in Figure 5, the results indicated that when basing the BPC score on all available BPC data, the correlation between BPC and TPOCS-S was moderate, r (62) = .26. However, when limiting the BPC to the sessions also coded with the TPOCS-S, and as illustrated in Figure 6, the correlation between the two was much larger, r (60) = .75. ***This provides evidence that when scored for the course of treatment, the BPC and TPOCS-S lead to consistent conclusions.*** As illustrated in the figure, the majority of the deviations from the identity line reflect higher scores on the BPC. This indicates that for the course of treatment, therapists had a tendency to over-report the implementation of TF-CBT relative to the ratings by trained coders on the TPOCS-S. Likewise, if utilizing the BPC for ongoing TF-CBT fidelity monitoring, it is essential to provide periodic audits of therapists' reports against more objective evaluation methods.

Exploratory Aim 1. Test in a preliminary way the adequacy of the observational instrument for sessions conducted in Spanish as well as those conducted in English.

The TPOCS-S measurement models for session-by-session data were re-estimated using the subset of sessions conducted in Spanish and rated by one of two bilingual raters (who also rated sessions in English). This included 64 sessions, and due to the limited sample size, the analyses were only performed on the session-by-session data. The results are largely consistent with the TPOCS-S results for sessions conducted in English. First analyzed with the original 6-point rating scale for both the English and Spanish sessions, the dimensionality results were consistent, with variance explained of 33.0 (English) versus 28.0 (Spanish), eigenvalue for the first contrast of 2.1 for English and Spanish, and unexplained variance attributable to the first contrast of 11.8% versus 12.9%. The item fit statistics were also consistent. For the Spanish sessions, 2 items were significantly misfitting, and for the English sessions, 3 items were significantly misfitting. The most significantly misfitting item was the same for the English and Spanish sessions, Cognitive Coping (C.07E), with a value of 5.3 for English sessions and 2.5 for Spanish sessions. The performance of the rating scale was also highly consistent, with utilization of 64% versus 60% for category 1, 9% versus 10% for category 2, 8% v. 10% for category 3, 7% v. 7% for category 4, 6% v. 7% for category 5, 5% v. 5% for category 6, and 1% v. 1% for category 7. Likewise, the reliability and separation reliability statistics were equally low. Finally, the item-person map (Figure 7) illustrates that the session and item distributions were consistent, importantly, with the ordering of the TF-CBT PRACTICE elements from least to most difficult being similar to that observed for the English sessions (Figure 3). To further evaluate the performance of the TPOCS-S for English versus Spanish sessions, the Rasch measurement model was re-specified to include tests for differential item functioning (DIF). DIF reflects items that perform differently for subsets of the sample. For measurement, this is undesirable. DIF analysis is challenging because there is almost always some non-zero difference in the performance of items across sub-samples; therefore, it is important to determine how much DIF is required before remedies are implemented. If an item is identified that performs markedly differently in English and Spanish sessions, the most readily implemented remedy is to treat it as two separate items – an English version and a Spanish version. The DIF results, despite the modest sample size of coded Spanish sessions, were promising. The most dramatic DIF was observed for Assessment (C.01), which was less likely to occur in Spanish sessions. Given the modest number of Spanish sessions, this may reflect actual differences in the occurrence of this component between the two samples. There was also a significant difference for Parenting Skills (C.04), which was more likely to occur in Spanish sessions, and C.09 (Trauma Processing), which was less likely to occur in Spanish sessions. The evaluators suspect that the difference for Parenting Skills is attributable to the higher frequency of Spanish sessions that are parent or conjoint sessions, and that the difference for Trauma Processing is due to the modest sample of Spanish sessions. Anecdotally, the performance of the items across English and Spanish sessions was far more stable than is sometimes observed for other DIF analyses in fidelity measurement, and this is likely attributable to the quality of the coders and the coding protocol.

Exploratory Aim 2. Conduct a preliminary evaluation of relations among therapist variables and fidelity (Note: there too few cases per therapist to support a more robust evaluation).

To evaluate Exploratory Aim 2, the Rasch measurement model was applied to the combined English and Spanish TPOCS-S data. The resulting client-level scores for the course of TF-CBT were then specified as the outcome in a series of linear regression models that tested for differences in fidelity based on therapist characteristics, including: sex, ethnicity (Hispanic or Latino v. White v. Asian v. Other), job title (Intern and

Practicum v. Post-Doc and Psychologist v. Therapist and Senior Therapist v. Supervisor), time since hire at CII, first TF-CBT date since being hired (Less than 1 year v. 1 year or more), total TF-CBT clients since hire, TF-CBT clients in the study, time since training in TF-CBT (Less than 1 year v. 1 year or more), type of degree program (Master's v. Doctorate), years practicing in the field (recoded as 0, 1-5, 6-10, 11 or more), TF-CBT supervisor (Supervisor 1 v. Supervisor 2), Primary Supervisor (*n* = 19), total caseload during the study period, total caseload of TF-CBT clients during the study period, and program type (school-based v. outpatient v. intensive). Prior to analysis, the TPOCS-S Rasch score was modeled in a two-level mixed-effects regression model to determine the proportion of variance attributable to clients and therapists. The results indicated that nearly all (i.e., 99%) of the variance was attributable to clients, indicating that there was no nesting effect within therapists. Effectively, this means that no variance in TPOCS-S fidelity is available for explanation by therapist characteristics. Because there was no nesting effect, the client scores were treated as independent cases, and each group of therapist variables was evaluated in a separate regression model implemented through the general linear model in SPSS. Across models, two significant effects were detected. Male therapists had significantly lower TF-CBT extensiveness scores for the course of treatment, B = -1.22, SE = .49, $t(1)$ = -2.50, p = .015, and therapists who had a specific primary supervisor (i.e., one of the 19 supervisors) had significantly higher extensiveness scores, B = 1.12, SE = .54, $t(1)$ = 2.07, p = .043. When all therapist variables were combined in a simultaneous regression model, there were no statistically significant effects.

The same models were performed for the BPC. In contrast, the unconditional two-level mixed-effects model (clients nested within therapists) indicated that 19% of the variance was attributable to the therapist. As noted previously, the strong nesting effect likely reflects rating style of therapists providing self-reports. To evaluate therapist characteristics as predictors of fidelity, a series of two-level models were performed, with each group of therapist predictors entered at therapist level of the model. The results indicated that the 6 therapists with a Doctoral degree reported significantly lower levels of fidelity for the course of TF-CBT, γ_{01} = -1.53, SE = 0.64, $T(30)$ = -2.39, p = .02. Related to this, therapists with the title of Psychologist or Post-Doc Psychologist reported significantly lower levels of fidelity, γ_{02} = -1.62, SE = 0.64, $T(28)$ = -2.51, p = .02. Finally, therapists receiving TF-CBT supervision from one of the two supervisors reported significantly higher levels of fidelity, γ_{01} = 1.17, SE = 0.43, $T(30)$ = 2.69, p = .01. Follow-up analyses were conducted to evaluate possible explanations for this finding. The results indicated that the composition of the groups supervised by the two TF-CBT supervisors was quite different, with the supervisor having the higher BPC scores supervising more interns and all of the therapists in the Intensive program. When controlling this effect for other therapist variables (specifically, License Type, and whether the therapist was an Intern, Psychologist, Associate Clinical Social Worker, or LCSW or LMFT) the effect became non-significant.

Summary and Conclusions

Summary

The results of the current evaluation can be used to help answer the question, "To what extent did clients of CII receive TF-CBT as intended" over the course of treatment? This focus on what the client received is consistent with the measurement and reporting of adherence in psychotherapy efficacy and effectiveness trials, in which the key question is the extent to which observed differences in client outcomes are attributable to the experimental treatment having been delivered as intended with each client. Similarly, measurement methods used to evaluate therapist adherence in studies of the community-based implementation and outcomes of evidence-based treatments, although relatively few in number, also focus on the extent to which clients receive the treatments as intended.

Strengths of this study were the use and evaluation of two different methods to assess adherence to TF-CBT among a community-based sample of mental health clinicians: observational coding of audio-recorded treatment sessions (“gold standard”); and, a potentially feasible and low cost therapist self-report method. Data obtained using each of these methods were evaluated using measurement models grounded in Item Response Theory (IRT), specifically Rasch models. Such models produce a richer array of information about the nature of what is being measured and quality of measurement relative to traditional measurement models grounded in Classical Test Theory. The strengths and limitations of the observational and self-report measurement methods to assess adherence to TF-CBT at CII are summarized, as are important implications for the use of each method to support the implementation of TF-CBT in community-based practice.

The study evaluated adherence to TF-CBT as implemented at CII by 31 therapists with 94 clients between November 2013 and February 2015. Based on available therapist self-report data (i.e., the Brief Practice Checklist) that provided dates of coded treatment sessions, the average duration of TF-CBT treatment episodes was 18.79 weeks (*SD* = 7.66), during which time clients participated in an average of 15.75 treatment sessions (*SD* = 5.08). Of these sessions, approximately two-thirds (63.4%) were with children only, one-third (35%), with caregivers only, and only 1.6% were conjoint sessions. It is important to note that the prescribed protocol for TF-CBT specifies parallel child and caregiver sessions on an approximate weekly basis, with at least two of these as conjoint sessions to share the Trauma Narrative and address the Enhancing Safety component. As such, one might reasonably have expected a more even distribution of child and caregiver sessions than was found in the current sample, as well as a higher percentage of conjoint sessions. However, as noted in previous research (e.g., Cohen et al., 2011; Deblinger et al., 2011) and two recent reviews (Cary & McMillen, 2012; de Arellano et al., 2014), caregiver participation is a frequent challenge to TF-CBT implementation, which provides a plausible explanation for these findings.

TF-CBT TPOCS-S (TPOCS-S). The first level of analyses of the observational data indicated that the coder training procedures and coding protocol were implemented as intended, providing confidence in the obtained results, which are summarized briefly here.

What did it measure? Across a wide variety of analyses performed, there was consistent evidence that the TPOCS-S more accurately captured adherence to the TF-CBT components (the acronym for which is PRACTICE) across the *course* of treatment, rather than at the level of individual sessions. One of the primary reasons for this was poor targeting of items to individual sessions, namely that most items were unlikely to occur in a given session. These findings are consistent with the fact that the core components of TF-CBT are delivered over the course of treatment with different components addressed during each session. In other words, based on an estimated 18 treatment sessions, approximately one-third (sessions 1-6) are devoted to the PRAC components; one-third to TI (sessions 7-12); and one-third to CE (sessions 13-18). Thus, as a components-based model, it is expected that only one or two components would be addressed within a given session, making a fidelity measurement over the course of treatment the most meaningful.

Another finding regarding the TPOCS-S was that most of the variance (71%) was attributable to the items, with approximately one-third (29%) attributable to clients, and a very small percentage (<1%) to therapists; similar results were obtained for the reliability analyses (client-level reliability was relatively high, .78. Therapist-level reliability was very low, <1%). Taken together, these results indicate that adherence to TF-CBT as measured by the TPOCS-S reflects the treatment received by each client, and that scores for clients treated by the same therapist are somewhat independent. These findings suggest that clients seen by the same therapist have somewhat different experiences of TF-CBT. To the extent that TF-CBT is intended to be tailored in accordance with the symptoms and progress of each client, this variability of adherence among

clients treated by the same therapist is to be expected. And, adherence as measured by the TPOCS-S cannot be considered to reflect a stable attribute of a therapist, such as overall or average therapist adherence.

What did TF-CBT look like for CII clients? In terms of the likelihood of a client experiencing the PRACTICE components over the course of treatment and the extensiveness of their implementation, analysis of the TPOCS-S data indicated that clients with an average course of TF-CBT had at least a 50% chance of experiencing the following TF-CBT components: Cognitive Coping, Relaxation, Affective Expression and Modulation, Psychoeducation and Trauma Narrative. The finding that 84% of clients who experienced an average course of TF-CBT received the Trauma Narrative is particularly interesting given previous research, which indicated that community-based therapists are less likely to use trauma-specific techniques (Allen & Johnson; Jensen-Doss et al., 2008) and that the greatest perceived challenges to delivery of TF-CBT involved the Trauma Narrative (or gradual exposure) components (Hanson et al., 2012). However, while development of the Trauma Narrative had high extensive ratings, the Trauma Processing component was a low occurring item; a finding that is consistent with these prior research studies. In terms of adherence levels, the components with the highest level of extensiveness for an average client were Psychoeducation, Relaxation, Affective Expression, and the Trauma Narrative; and again, the specific Trauma Processing component emerged as a low frequency item, which is consistent with prior research and may reflect the challenges associated with its implementation.

Therapist self-reports on the Behavior Practice Checklist (BPC). Data from therapist self-reports of adherence in each treatment session as indexed by the BPC were also evaluated. When modeled on a session-by-session basis, the performance of the BPC was generally consistent with the TPOCS-S, thus the decision was made to similarly evaluate this measure over a course of treatment, rather than on a session-by-session basis. In terms of components implemented, data from the BPC indicated that an average client experienced approximately 28 of the 31 items indexing the PRACTICE components. Addressing cognitive distortions throughout the Trauma Narrative emerged as one of the less frequently occurring items, which is consistent with prior research. Over the course of treatment, the ordering of the treatment components largely mapped on to the PRACTICE components reflecting a progression from the easier to more complex components. That almost all items indexing the PRACTICE components are reported by therapists to occur is noteworthy, as described next.

Associations among TPOCS-S and BPC. For single sessions, the TPOCS-S generally reflected higher levels of implementation of TF-CBT relative to the BPC, but this is likely due to the different scale ranges (3-point extensiveness of use vs. dichotomous scale reflecting occurrence, respectively). In contrast, consistent findings were obtained when comparing the two measurement methods over a full course of treatment. However, it is important to note that, relative to observational data, therapists did over-report their use of TF-CBT components; a finding that attenuates somewhat the confidence in reliance on therapist self-report as an accurate reflection of session content. More specifically, when comparing reliability indices between the TPOCS-S and BPC, a higher proportion of variance was attributed to therapists on the BPC as compared to the TPOCS-S (56% vs. <1%). Thus, when the therapist provides ratings, rather an external, objective and trained rater, all of the information is filtered through that therapist's perspective. For an external rater of the same course of treatment, the therapist-level interrater reliability likely would be much lower, as was observed for the TPOCS-S.

Preliminary evaluation of the TPOCS-S for Spanish vs. English language sessions. To our knowledge, this is the first time this type of comparison has been conducted and findings are promising. In contrast to the other analyses, the evaluation of the Spanish TPOCS-S focused on session-by-session rather than treatment episode data owing to the limited number of sessions in the sample. Despite the limited available sample size, the ordering of TF-CBT PRACTICE components from least to most difficult was similar across the English and Spanish rating scales. A few differences were noted, in that Assessment and Trauma Processing were less likely to occur in Spanish sessions, whereas Parenting was more likely to occur. One plausible explanation for these findings is the higher frequency of Spanish sessions that were caregiver or conjoint and the modest sample of overall Spanish sessions which could account for the lower number of sessions devoted to the Trauma Narrative and Trauma Processing.

Associations among therapist variables and fidelity. Overall, few specific therapist factors appear to impact adherence to TF-CBT as measured observationally or via therapist self-report. Only two therapist variables emerged as factors associated with fidelity as measured by the TPOCS-S, with male therapists and those with one of the 19 primary supervisors having lower and higher (respectively) TF-CBT extensiveness

scores for the course of treatment. With respect to the BPC, it is important to bear in mind the strong nesting effects (19% of variance accounted for by therapist) that likely reflect the rating style of therapists in completing the checklists. Findings showed that therapists' level of educational degree and professional title were associated with lower levels of adherence. Therapists having a Doctoral degree and those with the title of Psychologist or Post-Doc Psychologist had significantly lower adherence scores. In addition, therapists receiving TF-CBT supervision from one supervisor had significantly higher adherence scores relative to therapists receiving TF-CBT supervision from the other supervisor. Follow-up analyses conducted to evaluate possible explanations indicated the composition of therapists in the two supervision groups varied systematically. There were more interns, and all therapists in the Intensive treatment program, in the group with the higher levels of adherence scores. When controlling for other therapist variables (specifically, License Type, and whether the therapist was an Intern, Psychologist, Associate Clinical Social Worker, or LCSW or LMFT) the supervisor effect became non-significant.

Conclusions

A primary aim of the evaluation was to examine the fidelity of TF-CBT implementation at CII. Results of analyses of the data obtained using the TPOCS-S reflected adherence to the TF-CBT PRACTICE components over the course of treatment experienced by a client. That is, the scores from this instrument accurately reflect the occurrence of TF-CBT PRACTICE components and extensiveness of their implementation over the course of a client's treatment. This information is useful in answering the adherence question, *"To what extent did this particular client get the treatment as intended?"* However, use of the TPOCS-S in community practice may be prohibitively expensive. In contrast, data obtained easily and cost effectively via the BPC are influenced significantly by the perception of the therapist, such that scores are likely to be more similar across clients treated by the same therapist than by clients treated by different therapists. Relative to observational data, therapists have a tendency to over-report their use of TF-CBT components. In addition, the BPC reflects only whether a component occurred, not the extensiveness with which it was implemented. Accordingly, BPC scores will provide less accurate answers to the question, *"To what extent did this particular client experience the treatment as intended,"* because the answer is significantly influenced by the perception or rating style of the particular therapist providing TF-CBT and because extensiveness of implementation is not indexed.

Results from the TPOCS-S (which represents the 'gold standard' for fidelity measurement) regarding the extent to which clients at CII experienced TF-CBT as intended, indicated that clients with an average course of TF-CBT had at least a 50% chance of experiencing five of the core components of TF-CBT (Psychoeducation, Relaxation, Affective Expression and Modulation, Cognitive Coping, and Trauma Narrative). These findings are consistent with prior research (e.g., Allen & Johnson, 2012; Jensen-Doss et al., 2008) indicating that while therapists report use of TF-CBT, they do not appear to be delivering the full, prescribed course of treatment. While this indicates problems with adherence, as noted previously, these results are in line with previous research demonstrating variations in fidelity across other evidence-based treatment interventions and the important role of contextual factors, such as organizational culture and climate which can significantly impact availability of ongoing training and supervision known to increase treatment fidelity (Beidas, Edmunds, Marcus, & Kendall 2012; McHugh & Barlow, 2010; Woody, Anderson & D'Souza, 2015).

On an encouraging note, the association between the BPC and TPOCS-S suggests that the two scores lead to similar conclusions. However, the nature of the "disagreements" between the two scores is informative – in the majority of cases, the score is higher for the BPC relative to the TPOCS-S, again likely due to the fact that therapists over-report their use of components, a limitation that must be acknowledged with its continued use. Appropriate use of the TPOCS-S and BPC are to index adherence to TF-CBT over the course of treatment for a client. TPOCS-S scores could be used in clinical supervision to review, upon case closure, which PRACTICE components the therapist used and how extensively. As the number of clients completing treatment increases, patterns of component use and their extensiveness can be considered in relationship to objectively measured treatment outcomes. With respect to the BPC, component use, but not extensiveness, could be observed and discussed in supervision and in relation to such client outcomes.

With these caveats in mind, the findings of this evaluation suggest that the BPC may provide an efficient, feasible and low-cost way to summarize adherence to TF-CBT for the course of a given client's treatment (as opposed to on a session-by-session basis). For example, BPC data from cases completed successfully (i.e. by mutual agreement of therapists and families, with evidence the intended outcomes were achieved) can be reviewed in supervision and compared with BPC data provided by the same therapist for a

case in which case closure was unplanned, and/or outcomes were not achieved. As the number of clients completing treatment increases, patterns of component use can be considered in relationship to objective indicators of treatment outcomes.

In addition, the session information reported by therapists on the BPC could be used to monitor whether the planned use of components in a specific treatment session discussed in supervision actually occurred. Note, however, because therapists over-report use of components relative to observational data, some circumspection would be warranted with respect to using session-by-session data in this way. Because therapist rating style and apparent over-endorsement of components are limitations of the BPC, routine audits of treatment fidelity, including periodic direct observation or review of audio-recorded treatment sessions, and of objective measures of client outcomes will need to continue to be core priorities of TF-CBT implementation.

Importantly, given the evidence from this evaluation that adherence differs among clients treated by the same therapist, effective methods to train and provide clinical support to therapists are likely to require simulated and real opportunities for the therapist to practice and receive feedback regarding implementation of the TF-CBT PRACTICE components across multiple clients. Ongoing clinical supervision of TF-CBT cases provides such opportunities.

Finally, the IRT-based measurement models used to evaluate the TPOCS-S and BPC are well suited to the nature of specification of TF-CBT via components for which the order of implementation is not tightly constrained and all components are not used in each session. This built-in variability in the use of TF-CBT treatment components is indexed accurately using Rasch models, in which the probability of specific treatment components occurring is captured. The probability data also reflect the difficulty of use of components. Neither type of information is captured in measurement models based in Classical Test Theory.

References

Aarons, G. A., Hurlburt, M., & Horwitz, S. M. (2011). Advancing a conceptual model of evidence-based practice implementation in public service sectors. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 4-23.

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Allen, B., & Johnson, J.C. (2012). Utilization and implementation of trauma-focused cognitive-behavioral therapy for the treatment of maltreated children. *Child Maltreatment, 17*(1), 80-85. doi: 10.1177/1077559511418220

Bearman , S.K., Weisz, J.R., Chorpita, B.F., Hoagwood, K. Ward, Ugueto, A.M., Bernstein, A., The Research Network on Youth Mental Health (2013). More practice, less preach? The role of supervision processes and therapist characteristics in EBP implementation. *Administration and Policy in Mental Health and Mental Health Services Research, 40:* 518-529. doi: 10.1007s/10488-013-0485-5.

Beidas, R.S., Edmunds, J.E., Marcus, S.C., & Kendall, P.C. (2012). Training and consultation to promote implementation of an empirically supported treatment: A randomized trial. *Psychiatric Services, 63,* 660-665. doi: 10.1176/appi.ps.201100401

Beretvas, S. N., & Kamata, A. (Eds.). (2005). The multilevel measurement model. [Special issue]. *Journal of Applied Measurement, 6*(3).

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Boxmeyer, C.L., Lochman, J.E., Powell, N.R., Windle, M.,& Wells, K. (2008, Fall). School counselors' implementation of Coping Power in a dissemination field trial: Delineating the range of flexibility with in fidelity. *Emotional and Behavioral Disorders in Youth,* pp. 79-84, 94-95.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO for Windows [Computer software]. Lincolnwood, IL: Scientific Software International.

Cary, C.E. & McMillen, C.J. (2012). The data behind dissemination: a systematic review of trauma-focused cognitive behavioral therapy for use with children and youth. *Children and Youth Services Review, 34,* 748–757.

Caroll, K.M., & Rounsaville, B.J. (2010). No train, no gain? *Clinical Psychology: Science and Practice, 17* (1), 36-40.

Cohen, J. A., Mannarino, A. P., & Deblinger, E. (2006). *Treating trauma and traumatic grief in children and adolescents.* New York, NY: Guilford Press.

Cohen, J. A., Mannarino, A. P., & Deblinger, E. (2012). *Trauma-focused CBT for children and adolescents: Treatment applications.* New York, NY: Guilford Press.

Cohen, J.A., & Mannarino, A.P. (1996). A treatment outcome study for sexually abused preschool children: Initial findings. *Journal of the American Academy of Child & Adolescent Psychiatry, 35*, 42-50. doi: 10.1097/00004583-199601000-00011.

Cohen, J.A., Mannarino, A.P., & Iyengar, S. (2011). Community treatment of posttraumatic stress disorder for children exposed to intimate partner violence: A randomized controlled trial. *Archives of Pediatrics and Adolescent Medicine, 165,* 16–21.

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

de Arellano, M.A.R., Lyman, R.D., Jobe-Shields, L., George, P., Dougherty, R.H., Daniels, A.S., et al., (2014). Trauma-focused cognitive-behavioral therapy for children and adolescents: Assessing the evidence. *Psychiatric Services, 65,* 591-602.

Deblinger, E., Mannarino, A.P., Cohen, J.A. & Steer, R.A. (2006). A follow-up study of a multisite, randomized, controlled trial for children with sexual abuse-related PTSD symptoms. *Journal of the American Academy of Child Psychiatry, 45,* 1474-1484.

Deblinger, E., Mannarino, A.P., Cohen, J.A., Runyon, M.K., & Steer, R.A. (2011). Trauma-focused cognitive behavioral therapy for children: Impact of the trauma narrative and treatment length. *Depression and Anxiety, 28,* 67-75. doi: 10.1002/da.20744

Deblinger, E., Dorsey, S., Cooper, B., McLeod, B., & Garland, A.F. (2013). *Scoring manual for the TF-CBT version of the Therapy Process Observational Coding System for Child Psychotherapy – TF-CBT TPOCS-S. Unpublished manuscript.*

Deblinger, E., Cohen, J. A., Mannarino, A. P., Runyon, M. K., & Hanson, R. (2008). *PRACTICE Treatment Adherence Checklist Scoring Sheet – Short Version.* Unpublished instrument, University of Medicine and Dentistry of New Jersey - School of Osteopathic Medicine, Stratford, New Jersey.

Deci, P.A., Santos, A.B., Hiott, D.W., Schoenwald, S., & Dias, J.K. (1995). Dissemination of Assertive Community Treatment programs. *Psychiatric Services, 46,* 676-687.

Dorsey, S., Pullman, M.D., Deblinger, E., Berliner, L., Kerns, S.E., Thompson, K., Garland, A. (2013). Improving practice in community-based settings: A randomized trial of supervision -- study protocol. *Implementation Science, 8,* 89. doi: 10.1186/1748-5908-8-89.

Fixsen, D. L, Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature.* Tampa, FL: University of South Florida, Louis de la Parte Florida Mental Health Institute, The National Implementation Research Network (FMHI Publication #231).

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., et al. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31,* 4-19.

Hogue, A., Liddle, H.A., Rowe, C. (1996). Treatment adherence process research in family therapy: A rationale and some practical guidelines. *Psychotherapy, 33*: 332-345.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*, 328.

Hanson, R.F. Gros, K. S., Davidson, T., Barr, S., Cohen, J., Deblinger, E., Mannarino, A. P., & Ruggiero, K. J. (2014). National trainers' perspectives on challenges to implementation of an empirically supported mental health treatment. *Administration and Policy in Mental Health and Mental Health Services Research, 41*(4), 522-534. doi: 10.1007/s10488-013-0492-6. PMCID: PMC3758397.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3,* 85-106.

Linacre, J. M. (2014a). *FACETS Rasch measurement computer program*. Beaverton, OR: Winsteps.com.

Linacre, J. M. (2014b). *WINSTEPS Rasch measurement computer program*. Beaverton, OR: Winsteps.com.

McLeod, B. D., Southam‐Gerow, M. A., Tully, C. B., Rodríguez, A., & Smith, M. M. (2013). Making a case for treatment integrity as a psychosocial treatment quality indicator for youth mental health care. *Clinical Psychology: Science and Practice, 20*(1), 14-32.

McHugh, R. K., & Barlow, D. H. (2010). The dissemination and implementation of evidence-based psychological treatments: A review of current efforts. *American Psychologist, 73,* 73–84. doi:10.1037/a0018121.

Muthén, L. K., & Muthén, B. O. (2014). *Mplus (version 7)* [computer software and manual]. Los Angeles, CA: Muthén & Muthén.

Nadeem, E., Gleacher, A., & Beidas, R. S. (2013). Consultation as an implementation strategy for evidence-based practices across multiple contexts: Unpacking the black box. *Administration and Policy in Mental Health and Mental Health Services Research, 40*(6), 439-450.

Perepletchikova, F., Treat, T.A. & Kazdin, A.E. (2007). Treatment integrity in psychotherapy research: Analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology, 75*, 829-841.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA press.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2013). *HLM 7: Hierarchical linear & nonlinear modeling (version 7.00)* [Computer software & manual]. Lincolnwood, IL: Scientific Software International.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measurement. *Quality of Life Research, 16,* 19-31.

Pollio, E., McLean, M., Behl, L.E., & Deblinger, E. (2014). Trauma-focused cognitive behavioral therapy. In R. Reece, R.F. Hanson, & J. Sargent (Eds.) *Treatment of child abuse 2nd ed.* (pp. 31-38). Baltimore, MD: Johns Hopkins University Press.

Saunders, B.E., Berliner, L., & Hanson, R.F. (Eds.). (2004). *Child Physical and Sexual Abuse: Guidelines for Treatment (Revised Report: April 26, 2004).* Charleston, SC: National Crime Victims Research and Treatment Center.

Schoenwald, S.K.**,** Chapman, J.E., Sheidow, A.J., & Carter, R.E. (2009). Long-term youth criminal outcomes in MST transport: The impact of therapist adherence and organizational climate and structure*. Journal of Clinical Child and Adolescent Psychology, 38,* 91 – 105. PMCID: PMC2929913

Schoenwald, S.K., Sheidow, A.J., & Chapman, J.E. (2009). Clinical supervision in treatment transport: Effects on adherence and outcomes. *Journal of Consulting and Clinical Psychology, 77,* 410-421.PMCID: PMC2762701

Schoenwald, S.K., & Garland, A.F. (2013). A review of treatment adherence measurement methods. *Psychological Assessment, 25*, 146-156. doi: 10.1037/a0029715

Schoenwald, S.K., Garland, A.F., Chapman, J.E., Frazier, S.L., Sheidow, A.J., & Southam-Gerow, M.A. (2011). Toward the effective and efficient measurement of implementation fidelity. *Administration and Policy in Mental Health and Mental Health Services Research, 38*, 32-43.doi: 10.1007/s10488-010-0321-0

Schumacker, R. E., & Smith, E. V., Jr. (2007). Reliability: A Rasch perspective*. Educational and Psychological Measurement, 67*, 394-409.

Smith, E. V. Jr. (2000). Metric development and score reporting in Rasch measurement. *Journal of Applied Measurement, 1,* 303-326.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1,*199-218.

Wang, W. C., & Liu, C. Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement, 67,* 583-605.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wolfe, E. W., & Smith, E. V., Jr. (2007). Instrument development tools and activities for measure validation using Rasch models. *Journal of Applied Measurement, 8,* 97-123.

Woody, J.D., Anderson, D.K., & D'Souza, H.J. (2015) Dissemination of trauma-focused cognitive behavioral therapy with community practitioners: Focus on self-efficacy. *Journal of Evidence-Informed Social Work, 12,* 323-335. doi: 10.1080/15433714.2014.950128.

Wright B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Yeaton, W.H., Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49*, 156 – 167.

Table 1

TF-CBT TPOCS-S Treatment Content Components

Number	Variable	Item Text
1	C.01E	Assessment
2	C.02E	Other Topics/Crisis or Case Management
3	C.03E	Psychoeducation about Trauma/TF-CBT
4	C.04E	Parenting Skills
5	C.05E	Relaxation
6	C.06E	Affective Expression and Modulation Skills
7	C.07E	Cognitive Coping
8	C.08E	Trauma Narrative
9	C.09E	Trauma Processing
10	C.10E	In Vivo Exposure
11	C.11E	Conjoint - Prepared Parent/Child for Conjoint Session
12	C.12E	Enhancing Safety

Table 2

Percentage of Sessions with TPOCS-S components endorsed

		%
C.01E	Assessment	10.4
C.02E	Other Topics/Crisis or Case Management	98.9
C.03E	Psychoeducation about Trauma/TF-CBT	38.6
C.04E	Parenting Skills	7.4
C.05E	Relaxation	55.3
C.06E	Affective Expression and Modulation Skills	69.8
C.07E	Cognitive Coping	27.1
C.08E	Trauma narrative	55.3
C.09E	Trauma Processing	14.7
C.10E	In Vivo Exposure	4.1
C.11E	Conjoint - Prepared Parent/Child for Conjoint Session	23.7
C.12E	Enhancing Safety	22.1

Table 3

TPOCS Percentage of Clients Receiving Components and Average Extensiveness Ratings

Variable	Element	% Receiving	Mean Rating (*SD*)
C.01E	Assessment	47.8	0.51 (0.56)
C.02E	Other Topics	100.0	1.46 (0.50)
C.03E	Psychoeducation	94.0	1.42 (0.61)
C.04E	Parenting Skills	22.4	0.24 (0.46)
C.05E	Relaxation	89.6	1.25 (0.64)
C.06E	Affective Expression	94.0	1.36 (0.60)
C.07E	Cognitive Coping	70.1	1.04 (0.81)
C.08E	Trauma narrative	83.6	1.46 (0.77)
C.09E	Trauma Processing	46.3	0.55 (0.66)
C.10E	In Vivo Exposure	25.4	0.25 (0.44)
C.11E	Conjoint Session	70.1	0.78 (0.57)
C.12E	Enhancing Safety	67.2	0.76 (0.61)

Table 4

PCA of Rasch Residuals

Variable	Element	Loading
C.03E	Psychoeducation	0.63
C.05E	Relaxation	0.51
C.06E	Affective Expression	0.51
C.01E	Assessment	0.41
C.02E	Other Topics	0.13
C.09E	Trauma Processing	-0.62
C.08E	Trauma Narrative	-0.61
C.11E	Conjoint Session	-0.56
C.10E	In Vivo Exposure	-0.52
C.04E	Parenting Skills	-0.22
C.07E	Cognitive Coping	-0.17
C.12E	Enhancing Safety	-0.03

Table 5

TPOCS-S Outfit Statistics for the Course of TF-CBT

Variable	Element	Outfit (zstd)
C.01E	Assessment	1.60
C.02E	Other Topics	-0.40
C.03E	Psychoeducation	2.10
C.04E	Parenting Skills	-0.10
C.05E	Relaxation	0.50
C.06E	Affective Expression	-0.10
C.07E	Cognitive Coping	0.70
C.08E	Trauma Narrative	-0.10
C.09E	Trauma Processing	0.30
C.10E	In Vivo Exposure	-0.80
C.11E	Conjoint Session	-2.60
C.12E	Enhancing Safety	0.30

Table 6

Predicted probability of each TPOCS component being delivered at all or with average or high extensiveness to an average client

		Extensiveness		
		Any	Average	High
C.01E	Assessment	0.43	0.14	0.03
C.02E	Other Topics/Crisis or Case Management	0.96	0.84	0.52
C.03E	Psychoeducation about Trauma/TF-CBT	0.95	0.81	0.48
C.04E	Parenting Skills	0.17	0.04	0.01
C.05E	Relaxation	0.91	0.70	0.34
C.06E	Affective Expression and Modulation Skills	0.94	0.77	0.43
C.07E	Cognitive Coping	0.84	0.53	0.20
C.08E	Trauma Narrative	0.96	0.84	0.52
C.09E	Trauma Processing	0.47	0.16	0.04
C.10E	In Vivo Exposure	0.18	0.05	0.01
C.11E	Conjoint - Prepared Parent/Child for Conjoint Session	0.67	0.31	0.09
C.12E	Enhancing Safety	0.66	0.30	0.08

Table 7

Brief Practice Checklist Items

Domain	Number	Variable	Item Text
Psychoeducation	1	P1.1	Therapists provided psycho-education
	2	P1.2	Made normalizing and validating statements
	3	P1.3	Reviewed limits of confidentiality
	4	P1.4	Laid out components of TF-CBT, and the length of treatment time
	5	P1.5	Engaged family
Parenting	1	P2.1	Therapist provided parenting skills
Relaxation	1	R.1	Therapist explained the physiology of relaxation and instructed on methods of relaxation
Affective Expression	1	A.1	Therapist assisted the child in accurately identifying their feelings, and various ways of regulating their emotions
	2	A.2	Name variety of feelings
	3	A.3	Link feelings to situations
	4	A.4	Link feelings to body and/or facial expressions
	5	A.5	Understand intensity of certain feelings
	6	A.6	Develop another way to talk about feelings
Cognitive Coping	1	C1.1	Therapist reviewed the cognitive triangle.
	2	C1.2	Distinguish between thoughts, feelings and actions
	3	C1.3	Educate child on connection between thoughts, feelings and actions
	4	C1.4	Help the child generate alternative thoughts that are more accurate or helpful, in order to feel differently.
Trauma Narrative	1	T.1	Therapist worked on a trauma narrative with the child
	2	T.2	Introduction, Title page, Timeline/Table of contents
	3	T.3	Got the details of the traumatic events
	4	T.4	Ask about thoughts and feelings throughout the narrative
	5	T.5	Worked to modify cognitive distortions throughout the narrative
	6	T.6	Reviewed the trauma narrative at the beginning of the session
	7	T.7	Did final chapter on "what they've learned, how they grew…"
Trauma Processing	8	T.8	Read the trauma narrative to a caregiver/supportive adult
In Vivo Exposure	1	I.1	Therapist developed an in-vivo desensitization plan to resolve avoidant behaviors
Conjoint	1	C2.1	Conjoint child-parent session; sharing trauma narrative with parent/caregiver
	2	C2.2	Prepared the caregiver
	3	C2.3	Prepared the child
Enhancing Safety	1	E.1	Therapist addressed the child's sense of safety and developed a safety plan
	2	E.2	Therapist taught problem-solving skills and/or social skills as needed by the child

Table 8

Percentage of sessions with BPC items endorsed

	Item Text	%
P1.1	Therapists provided psycho-education	19%
P1.2	Made normalizing and validating statements	26%
P1.3	Reviewed limits of confidentiality	9%
P1.4	Laid out components of TF-CBT, and the length of treatment time	19%
P1.5	Engaged family	11%
P2.1	Therapist provided parenting skills	18%
R.1	Therapist explained the physiology of relaxation and instructed on methods of relaxation	25%
A.1	Therapist assisted the child in accurately identifying their feelings, and various ways of regulating their emotions	18%
A.2	Name variety of feelings	16%
A.3	Link feelings to situations	17%
A.4	Link feelings to body and/or facial expressions	12%
A.5	Understand intensity of certain feelings	11%
A.6	Develop another way to talk about feelings	9%
C1.1	Therapist reviewed the cognitive triangle.	14%
C1.2	Distinguish between thoughts, feelings and actions	14%
C1.3	Educate child on connection between thoughts, feelings and actions	12%
C1.4	Help the child generate alternative thoughts that are more accurate or helpful, in order to feel differently.	11%
T.1	Therapist worked on a trauma narrative with the child	23%
T.2	Introduction, Title page, Timeline/Table of contents	14%
T.3	Got the details of the traumatic events	17%
T.4	Ask about thoughts and feelings throughout the narrative	21%
T.5	Worked to modify cognitive distortions throughout the narrative	13%
T.6	Reviewed the trauma narrative at the beginning of the session	17%
T.7	Did final chapter on "what they've learned, how they grew…"	8%
T.8	Read the trauma narrative to a caregiver/supportive adult	5%
I.1	Therapist developed an in-vivo desensitization plan to resolve avoidant behaviors	7%
C2.1	Conjoint child-parent session; sharing trauma narrative with parent/caregiver	5%
C2.2	Prepared the caregiver	6%
C2.3	Prepared the child	6%
E.1	Therapist addressed the child's sense of safety and developed a safety plan	9%
E.2	Therapist taught problem-solving skills and/or social skills as needed by the child	11%

Table 9

PCA of Rasch Residuals

Element		Loading
T.4	Ask about thoughts and feelings throughout the narrative	0.70
T.6	Reviewed the trauma narrative at the beginning of the session	0.67
T.1	Therapist worked on a trauma narrative with the child	0.64
T.3	Got the details of the traumatic events	0.64
T.5	Worked to modify cognitive distortions throughout the narrative	0.60
T.2	Introduction, Title page, Timeline/Table of contents	0.42
T.7	Did final chapter on "what they've learned, how they grew…"	0.42
T.8	Read the trauma narrative to a caregiver/supportive adult	0.13
I.1	Therapist developed an in-vivo desensitization plan to resolve avoidant behaviors	0.12
C2.3	Prepared the child	0.10
C2.2	Prepared the caregiver	0.06
C2.1	Conjoint child-parent session; sharing trauma narrative with parent/caregiver	0.04
E.1	Therapist addressed the child's sense of safety and developed a safety plan	0.03
E.2	Therapist taught problem-solving skills and/or social skills as needed by the child	0.00
A.2	Name variety of feelings	-0.67
A.3	Link feelings to situations	-0.66
A.4	Link feelings to body and/or facial expressions	-0.66
A.1	Therapist assisted the child in accurately identifying their feelings, and various ways of regulating their emotions	-0.58
A.5	Understand intensity of certain feelings	-0.56
A.6	Develop another way to talk about feelings	-0.55
P1.1	Therapists provided psycho-education	-0.16
P1.2	Made normalizing and validating statements	-0.15
P1.4	Laid out components of TF-CBT, and the length of treatment time	-0.15
P2.1	Therapist provided parenting skills	-0.15
P1.3	Reviewed limits of confidentiality	-0.14
P1.5	Engaged family	-0.11
R.1	Therapist explained the physiology of relaxation and instructed on methods of relaxation	-0.09
C1.2	Distinguish between thoughts, feelings and actions	-0.09
C1.1	Therapist reviewed the cognitive triangle.	-0.06
C1.3	Educate child on connection between thoughts, feelings and actions	-0.06
C1.4	Help the child generate alternative thoughts that are more accurate or helpful, in order to feel differently.	-0.02

Table 10

Rasch Outfit Statistics

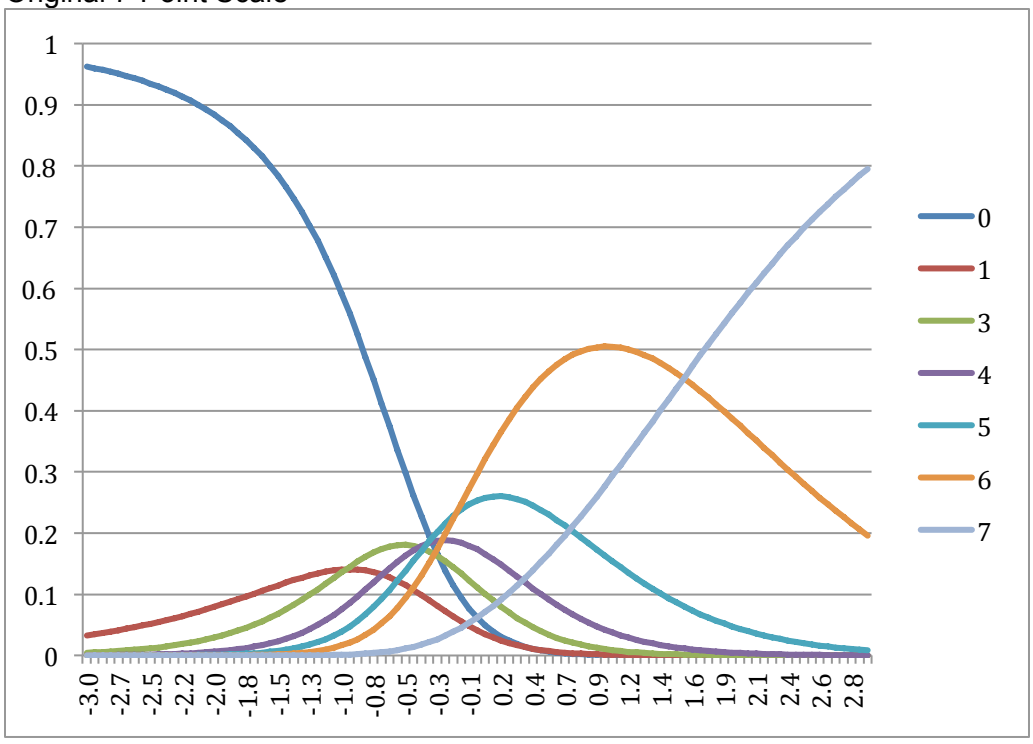
Variable	Element	Outfit (zstd)
P1.1	Therapists provided psycho-education	2.5
P1.2	Made normalizing and validating statements	3.6
P1.3	Reviewed limits of confidentiality	2.9
P1.4	Laid out components of TF-CBT, and the length of treatment time	1.3
P1.5	Engaged family	2.4
P2.1	Therapist provided parenting skills	2.2
R.1	Therapist explained the physiology of relaxation and instructed on methods of relaxation	-1
A.1	Therapist assisted the child in accurately identifying their feelings, and various ways of regulating their emotions	2.1
A.2	Name variety of feelings	-0.2
A.3	Link feelings to situations	-1.1
A.4	Link feelings to body and/or facial expressions	-0.8
A.5	Understand intensity of certain feelings	1.7
A.6	Develop another way to talk about feelings	1.1
C1.1	Therapist reviewed the cognitive triangle.	0
C1.2	Distinguish between thoughts, feelings and actions	-1.8
C1.3	Educate child on connection between thoughts, feelings and actions	-1.2
C1.4	Help the child generate alternative thoughts that are more accurate or helpful, in order to feel differently.	1.8
T.1	Therapist worked on a trauma narrative with the child	0.4
T.2	Introduction, Title page, Timeline/Table of contents	-2
T.3	Got the details of the traumatic events	-0.8
T.4	Ask about thoughts and feelings throughout the narrative	-1.9
T.5	Worked to modify cognitive distortions throughout the narrative	-0.1
T.6	Reviewed the trauma narrative at the beginning of the session	-1.7
T.7	Did final chapter on "what they've learned, how they grew…"	-1.5
T.8	Read the trauma narrative to a caregiver/supportive adult	-0.4
I.1	Therapist developed an in-vivo desensitization plan to resolve avoidant behaviors	-0.5
C2.1	Conjoint child-parent session; sharing trauma narrative with parent/caregiver	-1.2
C2.2	Prepared the caregiver	-0.9
C2.3	Prepared the child	-0.5
E.1	Therapist addressed the child's sense of safety and developed a safety plan	-0.6
E.2	Therapist taught problem-solving skills and/or social skills as needed by the child	-0.2

Table 11

Predicted probability of each BPC component being received by clients who experienced low, average, and high levels of adherence

		Level of Adherence		
		Low (-1 *SD*)	Average	High (+1 *SD*)
P1.1	Therapists provided psycho-education	0.99	1.00	1.00
P1.2	Made normalizing and validating statements	0.99	1.00	1.00
P1.3	Reviewed limits of confidentiality	0.61	0.92	0.99
P1.4	Laid out components of TF-CBT, and the length of treatment time	0.87	0.98	1.00
P1.5	Engaged family	0.61	0.92	0.99
P2.1	Therapist provided parenting skills	0.67	0.93	0.99
R.1	Therapist explained the physiology of relaxation and instructed on methods of relaxation	0.98	1.00	1.00
A.1	Therapist assisted the child in accurately identifying their feelings, and various ways of regulating their emotions	0.84	0.97	1.00
A.2	Name variety of feelings	0.91	0.99	1.00
A.3	Link feelings to situations	0.95	0.99	1.00
A.4	Link feelings to body and/or facial expressions	0.87	0.98	1.00
A.5	Understand intensity of certain feelings	0.73	0.95	0.99
A.6	Develop another way to talk about feelings	0.61	0.92	0.99
C1.1	Therapist reviewed the cognitive triangle.	0.70	0.94	0.99
C1.2	Distinguish between thoughts, feelings and actions	0.84	0.97	1.00
C1.3	Educate child on connection between thoughts, feelings and actions	0.78	0.96	0.99
C1.4	Help the child generate alternative thoughts that are more accurate or helpful, in order to feel differently.	0.73	0.95	0.99
T.1	Therapist worked on a trauma narrative with the child	0.56	0.90	0.98
T.2	Introduction, Title page, Timeline/Table of contents	0.76	0.96	0.99
T.3	Got the details of the traumatic events	0.67	0.93	0.99
T.4	Ask about thoughts and feelings throughout the narrative	0.78	0.96	0.99
T.5	Worked to modify cognitive distortions throughout the narrative	0.31	0.76	0.96
T.6	Reviewed the trauma narrative at the beginning of the session	0.67	0.93	0.99
T.7	Did final chapter on "what they've learned, how they grew…"	0.44	0.85	0.97
T.8	Read the trauma narrative to a caregiver/supportive adult	0.16	0.56	0.90
I.1	Therapist developed an in-vivo desensitization plan to resolve avoidant behaviors	0.18	0.61	0.92
C2.1	Conjoint child-parent session; sharing trauma narrative with parent/caregiver	0.23	0.68	0.94
C2.2	Prepared the caregiver	0.33	0.77	0.96
C2.3	Prepared the child	0.26	0.71	0.94
E.1	Therapist addressed the child's sense of safety and developed a safety plan	0.36	0.80	0.97
E.2	Therapist taught problem-solving skills and/or social skills as needed by the child	0.36	0.80	0.97

Original 7-Point Scale



3-Point Scale

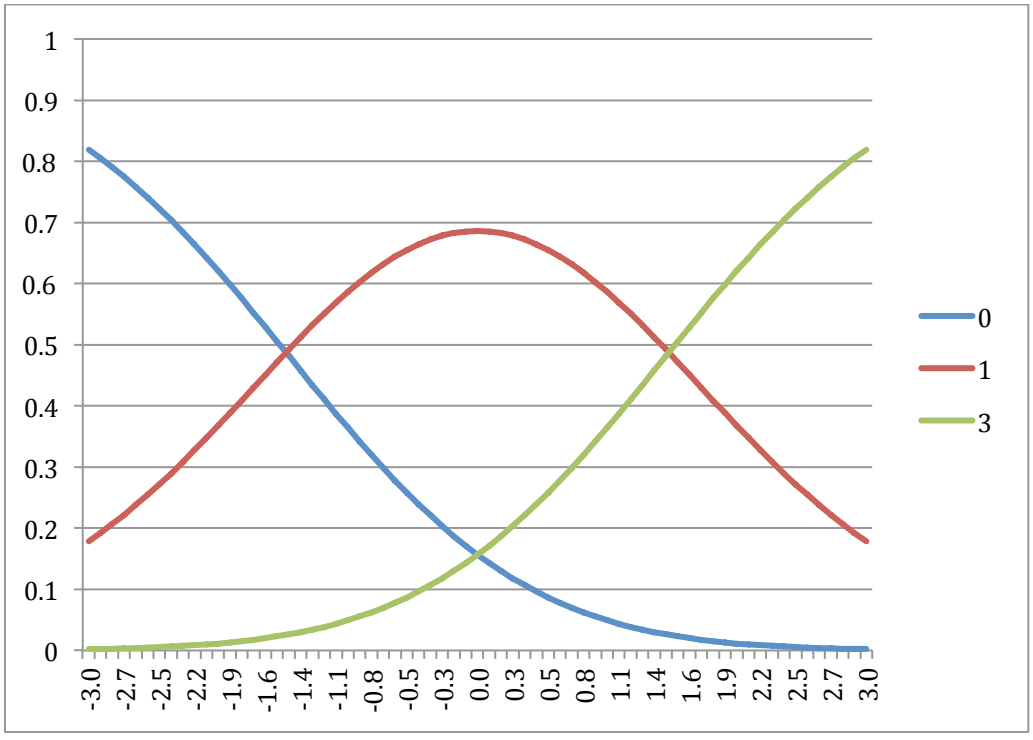


Figure 1. TPOCS-S Rating Scale Category Probability Curves, with the original 7-point rating scale and the optimized 3-point rating scale.

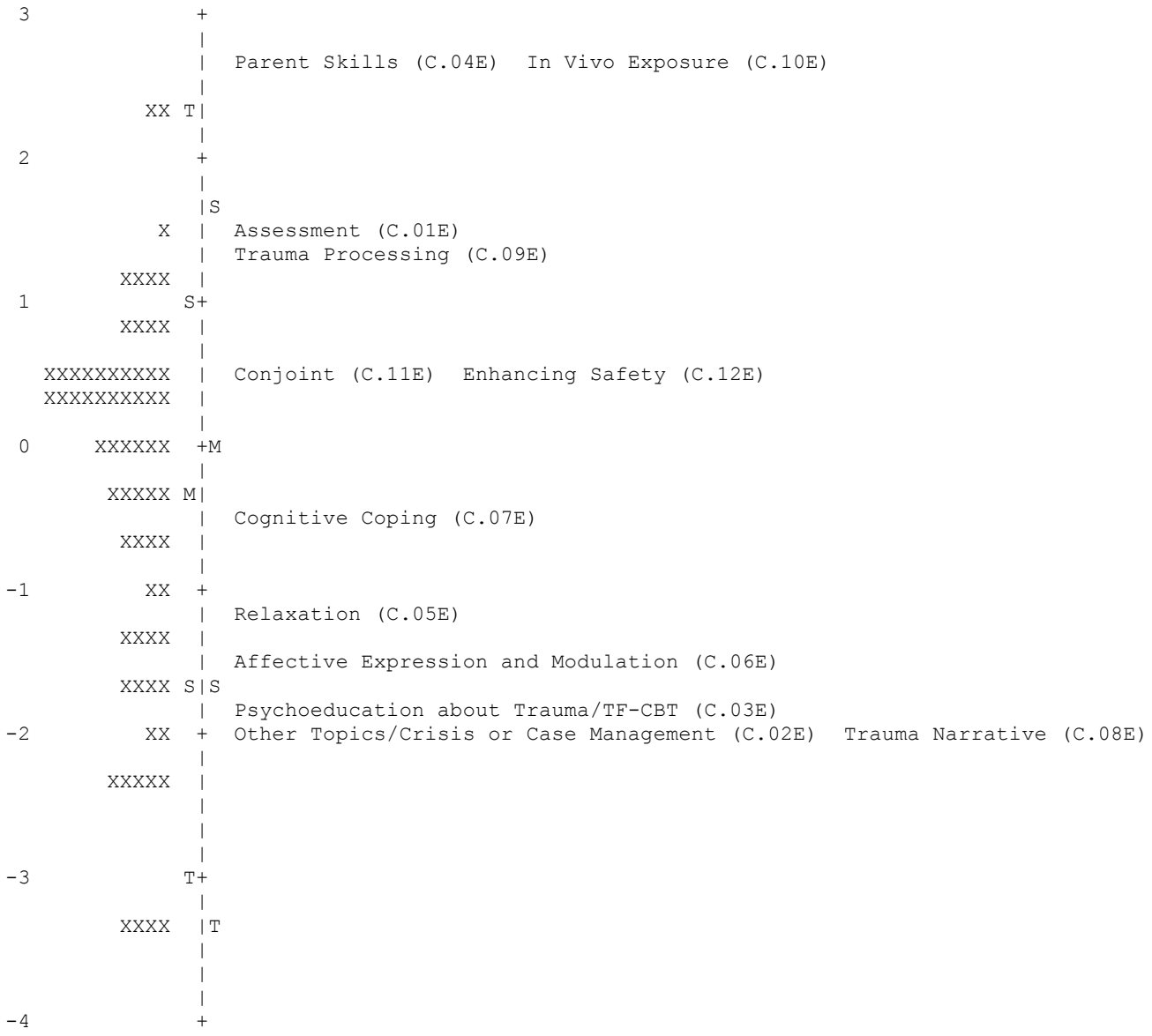


Figure 2. Rasch item-person map for the TPOCS-S for the course of TF-CBT for English sessions and using the 3-point rating scale.

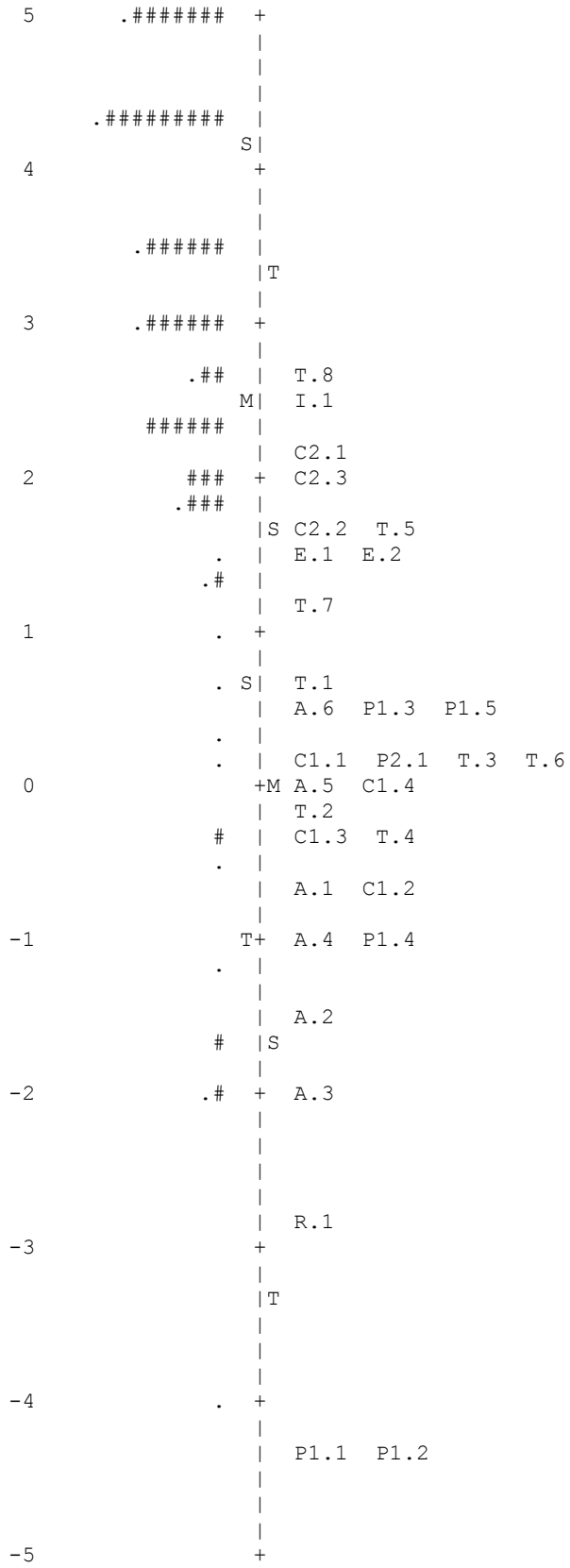


Figure 3. Rasch item-person map based on BPC data from the course of treatment, reflecting whether an item was ever endorsed.

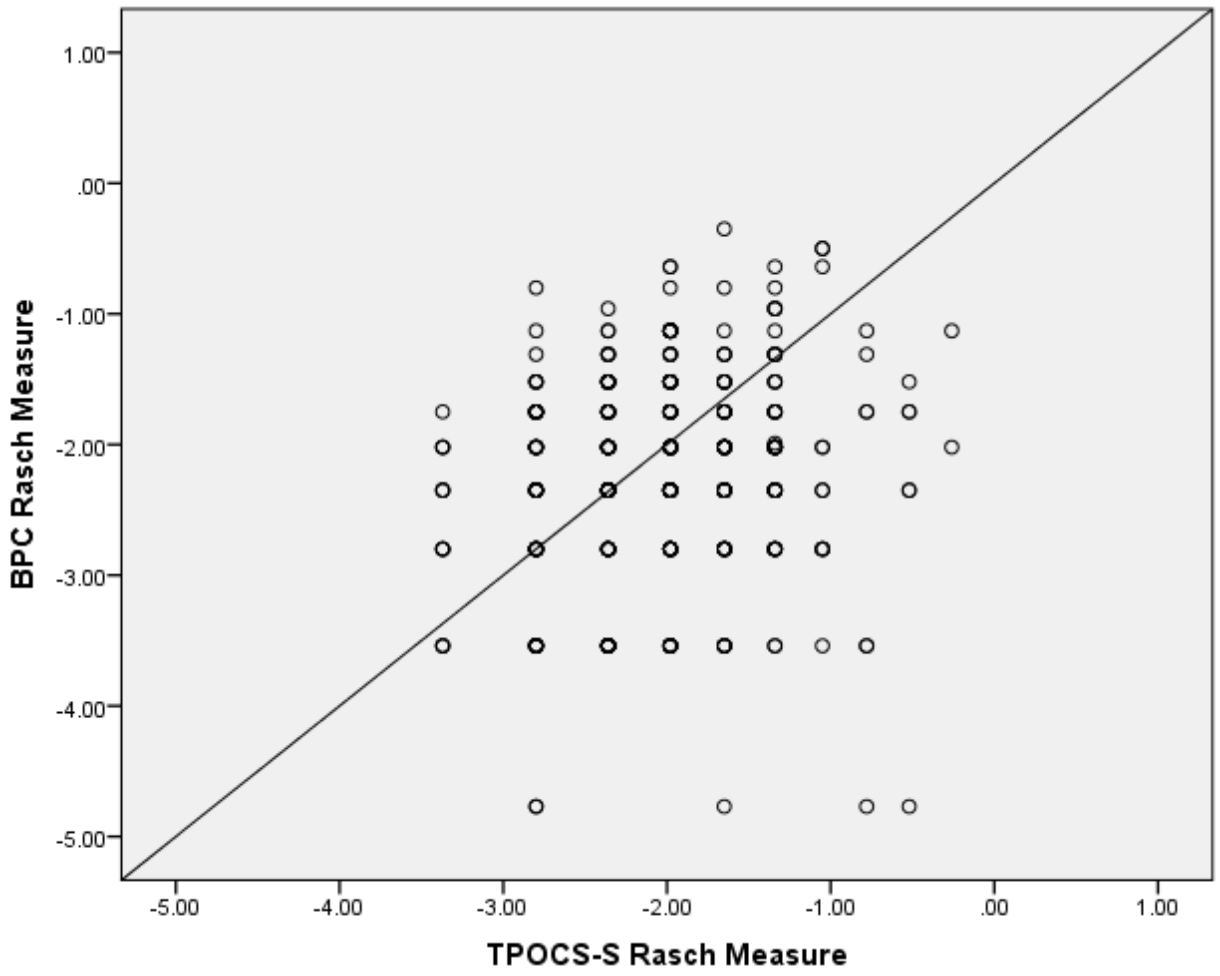


Figure 4. Scatterplot of Rasch scores for the TF-CBT TPOCS-S versus the BPC using session-by-session data.

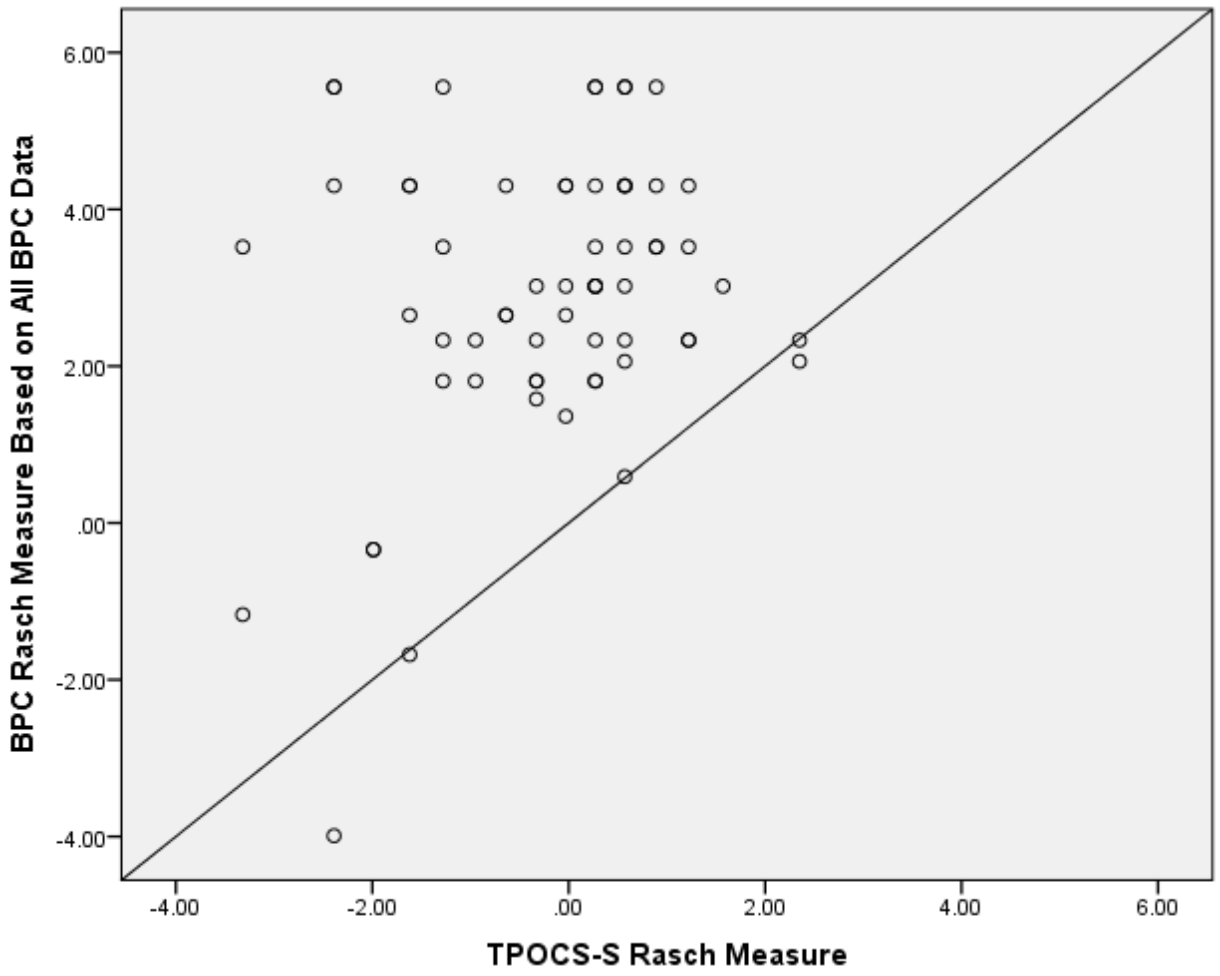


Figure 5. Scatterplot of Rasch scores for the course of TF-CBT. The score for the BPC is based on all available BPC data.

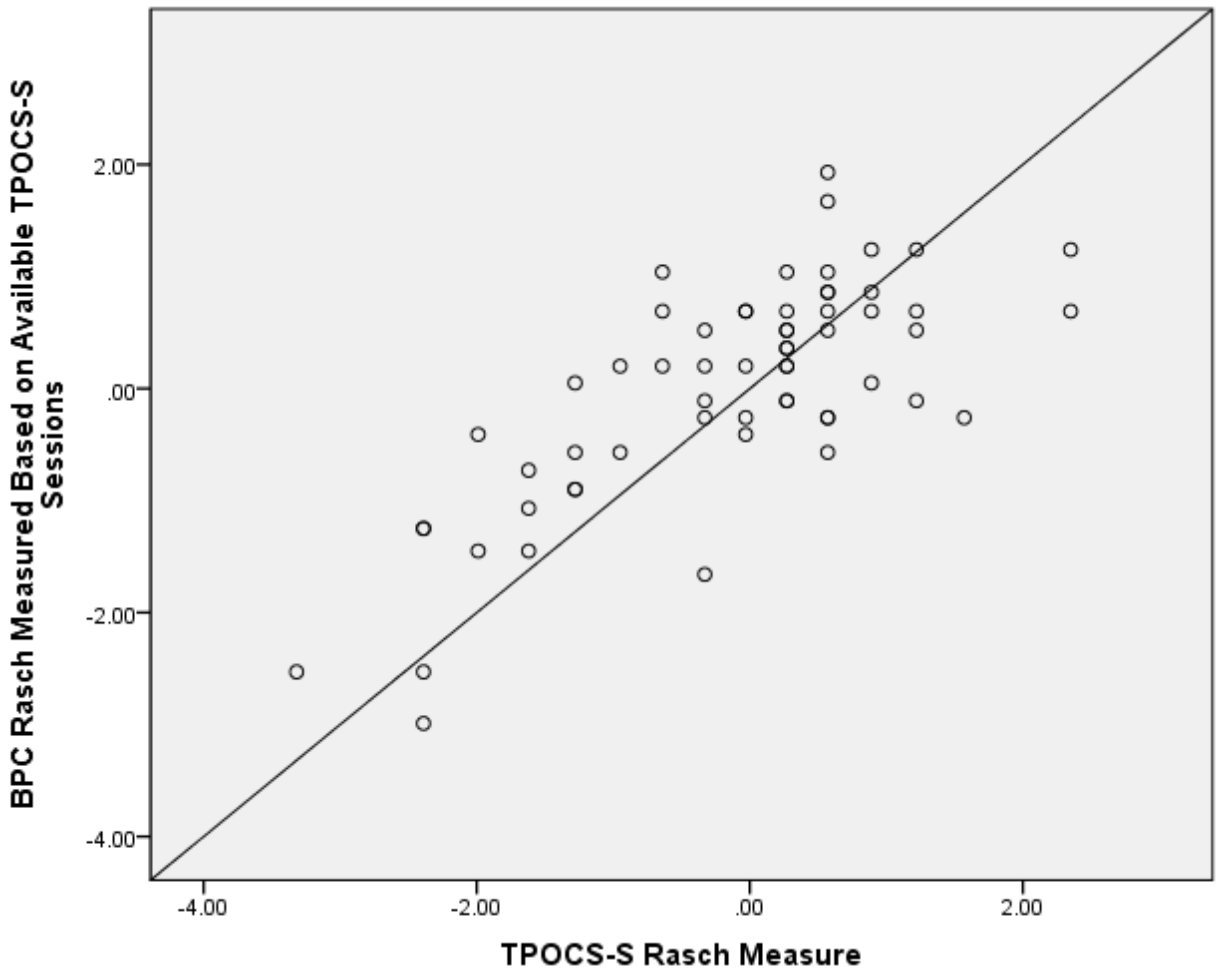


Figure 6. Scatterplot of Rasch scores for the course of TF-CBT. The score for the BPC is based on sessions that were also coded on the TPOCS-S.

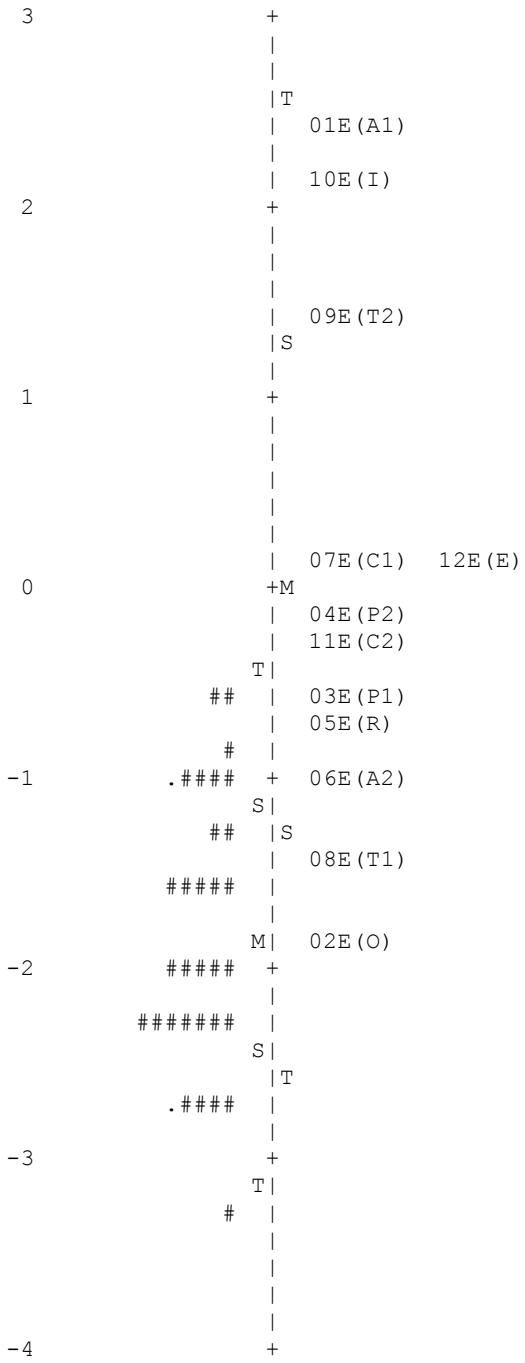


Figure 7. Session-person map for Spanish sessions rated with the TPOCS-S.

About MDRC

MDRC is a nonprofit, nonpartisan social and education policy research organization dedicated to learning what works to improve the well-being of low-income people. Through its research and the active communication of its findings, MDRC seeks to enhance the effectiveness of social and education policies and programs.

Founded in 1974 and located in New York City and Oakland, California, MDRC is best known for mounting rigorous, large-scale, real-world tests of new and existing policies and programs. Its projects are a mix of demonstrations (field tests of promising new program approaches) and evaluations of ongoing government and community initiatives. MDRC's staff bring an unusual combination of research and organizational experience to their work, providing expertise on the latest in qualitative and quantitative methods and on program design, development, implementation, and management. MDRC seeks to learn not just whether a program is effective but also how and why the program's effects occur. In addition, it tries to place each project's findings in the broader context of related research — in order to build knowledge about what works across the social and education policy fields. MDRC's findings, lessons, and best practices are proactively shared with a broad audience in the policy and practitioner community as well as with the general public and the media.

Over the years, MDRC has brought its unique approach to an ever-growing range of policy areas and target populations. Once known primarily for evaluations of state welfare-to-work programs, today MDRC is also studying public school reforms, employment programs for ex-offenders and people with disabilities, and programs to help low-income students succeed in college. MDRC's projects are organized into five areas:

- Promoting Family Well-Being and Children's Development
- Improving Public Education
- Raising Academic Achievement and Persistence in College
- Supporting Low-Wage Workers and Communities
- Overcoming Barriers to Employment

Working in almost every state, all of the nation's largest cities, and Canada and the United Kingdom, MDRC conducts its projects in partnership with national, state, and local governments, public school systems, community organizations, and numerous private philanthropies.